DOCUMENT RESUME

ED 421 535 TM 028 861

AUTHOR Parshall, Cynthia G.; Kromrey, Jeffrey D.; Chason, Walter M.; Yi, Qing

TITLE Evaluation of Parameter Estimation under Modified IRT Models and Small Samples.

PUB DATE 1997-06-00

NOTE 64p.; Paper presented at the Annual Meeting of the Psychometric Society (Gatlinburg, TN, June 26-29, 1997).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Ability; Error of Measurement; *Estimation (Mathematics); *Item Response Theory; Models; *Sample Size; Simulation; Tables (Data); *Test Items

IDENTIFIERS *Accuracy; Item Parameters

ABSTRACT

Accuracy of item parameter estimates is a critical concern for any application of item response theory (IRT). However, the necessary sample sizes are often difficult to obtain in practice, particularly for the more complex models. A promising avenue of research concerns modified item response models. This study both replicates and improves on an earlier investigation into modified models (C. Parshall, J. Kromrey, and W. Chason, 1996), which found tentatively positive results. To obtain realistic data, empirical item parameters were generated by fitting a six-dimensional model to archival data, using NOHARM (Fraser and McDonald, 1988). These parameters were then used along with thetas generated from independent normal ability distributions to generate simulated item response data. One hundred datasets were generated for each of four sample sizes. Finally, BILOG (Mislevy and Bock, 1990) was used to obtain estimated item ability parameters for each of the six investigated models. Results were evaluated in terms of accuracy and stability across samples. Accuracy was assessed as the degree to which both the obtained item responses and the known response probabilities were reproduced from the generating parameters. Stability was assessed as empirical estimates of standard errors. Crossvalidation of fit and accuracy was accomplished by applying the sample item parameter estimates to additional samples generated from the same population. (Contains 5 tables, 17 figures, and 27 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 535

Evaluation of Parameter Estimation under Modified IRT Models and Small Samples

Cynthia G. Parshall
Jeffrey D. Kromrey
Walter M. Chason
Qing Yi
University of South Florida

Abstract

Accuracy of item parameter estimates is a critical concern for any application of item response theory (IRT). However, the necessary sample sizes are often difficult to obtain in practice, particularly for the more complex models. A promising avenue of research concerns *modified* item response models. This study both replicates and improves upon an earlier investigation into modified models (Parshall, Kromrey, and Chason, 1996), which found tentatively positive results.

To obtain realistic data, empirical item parameters were generated by fitting a 6-dimensional model to archival data, using NOHARM. These parameters were then used along with thetas generated from independent normal ability distributions to generate simulated item response data. One hundred datasets were generated for each of four sample sizes. Finally, BILOG was used to obtain estimated item and ability parameters for each of the six investigated models.

Results were evaluated in terms of accuracy and stability across samples. Accuracy was assessed as the degree to which both the obtained item responses and the known response probabilities were reproduced from the generating parameters. Stability was assessed as empirical estimates of standard errors. Crossvalidation of fit and accuracy was accomplished by applying the sample item parameter estimates to additional samples generated from the same population.

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Cynthia Parshall

TM028861

Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN, June 26-29, 1997.

Small Samples and Modified Models:
An Investigation of IRT Parameter Recovery

The advantages of item response theory (IRT) for testing have been discussed theoretically for a number of years (Hambleton & Swaminathan, 1985; Lord, 1980). The benefits to testing programs include applications for test development, equating, and computer adaptive testing. The most popular models applied in practice are the unidimensional 1-parameter, 2-parameter, and 3-parameter logistic models (or, 1-PL, 2-PL, and 3-PL respectively). The formulas for these models are defined as

1-PL:

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

2-PL:

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

3-PL:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-a(\theta - b)}}$$

where

- $P(\theta)$ = the probability of a correct responses for an examinee of proficiency θ
- b = the item difficulty parameter
- a = the item discrimination parameter
- c = the lower asymptote, or item guessing parameter
- e = the base of the natural logarithms

These models all reflect the relationship between examinee proficiency, item characteristics, and response probabilities. The 1-parameter model has the single *b* parameter, or difficulty parameter. The 2-parameter model includes both the *b* and the *a* parameters. The *a* parameter allows individual discrimination values for each item (while the 1-parameter model constrains the items to the same discrimination value, allowing items to differ only in difficulty).

The 3-parameter model adds the *c* parameter, which approximates guessing, or chance-level responses above zero.

The number of item parameters which must be estimated in these models determines the examinee sample sizes required for calibrating the data. Although the recommendations for minimal sample size vary somewhat, typical guidelines are: 1000 examinees for the 3-parameter model, 500 examinees for the 2-parameter (Hulin, Lissak, & Drasgow, 1982), and 200 examinees for the 1-parameter model (Wright & Stone, 1979). The advantages of IRT methods will only be realized to the extent that the assumptions of the model utilized are met and that model-data fit is found. Among other possible problems, tests which are constructed based upon imprecise item parameters may result in an overestimate of the test information function and in ability estimates which are less accurate than they appear to be (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993). One source of poor parameter estimates is the use of an inadequate sample size for calibration, which can result in excessively large standard errors of the item parameter estimates (Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993; de Jong & Stoyanova, 1994).

Many testing programs are interested in using IRT methods for test development, analysis, and adaptive testing. However, the sample sizes required for stable parameter estimation are often not available in practice, particularly for the more complex models. The large sample sizes may be difficult to obtain if testing programs have small numbers of examinees per administration, if sub-group analyses draw from small numbers of examinees, if multiple sub-content areas are assessed separately, or if test forms are replaced frequently.

Sample size constraints might lead testing practitioners to select the model with the least stringent requirement (e.g., the one parameter). In practice, however, many testing programs use multiple choice items which vary in discrimination and allow for guessing. This would suggest that a more general model, such as the three parameter model, might provide the best fit to typical data and that use of a more limited model would lead instead to model misspecification errors (Divgi, 1986).

Spray, Kalohn, Schulz, and Fleer (1995) conducted a simulation to investigate the effect on adaptive classification testing when the true model was the 3-PL model, but the items were calibrated according to the 1-PL model. These researchers found use of the 1-PL model under

studied conditions to result in unacceptable rates of both false positive and false negative decisions (i.e., examinees classified as either passing and failing, who would have been classified otherwise according to the true 3-PL model). Yen (1981) has also pointed out problems which may arise when a 1-parameter or 2-parameter model is used inappropriately, or when truth is best modeled by a 3-parameter model. These problems include the potential for sample dependency of some item parameters, inaccurate model predictions, and attenuated correlations between actual and estimated trait values.

Modified Models

Given limitations on available examinee sample sizes, a practical concern is to obtain the most accurate item parameter estimates possible. A promising avenue of research concerns *modified* item response models (Barnes & Wise, 1991; Harwell & Janosky, 1991; Sireci, 1992; Stone & Lane, 1991). One type of model modification includes additional parameters in the model, while limiting estimation by fixing the value of the included parameters. Other modifications allow one or more included parameters a limited range within which they may vary.

Sireci (1992) investigated modifications to 1-PL and 2-PL models on multiple small sample datasets, obtained over several test administrations. Part of this study was an investigation into a modified model which included a fixed *c* parameter. One analysis considered restricted conditions, in which item parameters were constrained to be equal across the multiple samples of examinees. Another analysis addressed the use of mixed models (e.g., more than one IRT model for a specific analysis). Modified IRT models were also used by Stone and Lane (1991). In this study, an unconstrained 2-parameter IRT model was compared to a model in which item parameters were constrained to be equal across pretest-posttest administrations. This modification enabled an investigation into the stability of the item parameter estimates over time. Additional alternative IRT models have also been utilized in the context of differential item functioning (DIF) analysis (Thissen, Steinberg, & Wainer, 1993).

While some of the studies of modified item response models have been conducted on real data (Sireci, 1992; Stone & Lane, 1991), others have been simulations (Barnes & Wise, 1991;

Harwell & Janosky, 1991; Patsula & Pashley, 1996). Simulation studies have the advantage of utilizing true parameter values, which are never known in practice.

For example, Barnes and Wise (1991) conducted a simulation in which the item parameter estimates obtained under small sample conditions for typical 1-parameter and 3-parameter models were compared to two modified models. The modifications in this study involved the inclusion of a fixed, non-zero *c* parameter. These fixed *c* parameter models were based on the number of response options in the multiple choice items, *A*. One modification fixed *c* at $1/A$, and a second modification fixed *c* at $1/A$ - .05. Because the value of the *c* parameter was fixed, the sample size requirements for a standard 1-parameter model remained appropriate under the modifications. The results indicated that the modified models yielded more accurate parameter estimates than the more traditional 1-parameter and 3-parameter models.

Harwell and Janosky (1991) also investigated item parameter estimation with small samples. This simulation study examined several 2-parameter models in which estimation of the *a* parameter was affected by imposing different variances on the prior distribution of the *a*'s. Under the conditions in this study, item parameter estimates for small samples were recovered more accurately when a more informative (i.e., narrower) prior variance was used.

Parshall, Kromrey, and Chason (1996) also investigated the constrained *a* parameter approach, as well as an approach which utilized a fixed *c* parameter. This study utilized simulated data based upon parameters obtained from an achievement test of moderate length. The modified models examined showed some improvement in fit and stability, over unmodified models with the same number of parameters, under the studied conditions.

An alternative approach investigated by Patsula and Pashley (1996) used polynomial logistic regression to model ICCs in pretest items (i.e., when ability estimates can be reliably computed based on operational items). This procedure included a mixed model component in that it provided a means of identifying subsets of items which could be adequately modeled with fewer parameters (i.e., 2-PL or 1-PL). Where a reduced number of parameters needs to be estimated, presumably more stable results can be obtained under smaller sample conditions.

The results of all these studies suggest that modifications to popular IRT models are worthy of further investigation, and that appropriate modifications may provide more stable estimation of parameters with fewer examinees than unmodified models. This study was

intended to build upon previous research into modified item response models, under moderate and small sample size conditions. Secondly, this study included a greater number of replications than are often found in parameter estimation studies, providing for a more stable analysis of results (Robey & Barcikowski, 1992; Stone, 1992). Finally, this study used MIRT for data simulation. Data generated with this technique more closely approximates actual observed test data (Davey, Nering, & Thompson, 1997).

Educational Importance

Obtaining accurate parameter estimation is a critical concern, since all of the applications of IRT are based on these parameters. However practical testing applications frequently include elements which might best be modeled by more complex models (e.g., the 3-parameter model) while having only small samples of examinees to draw upon for calibration data. Determining a means for parameter estimation under these conditions provides the opportunity to make correct application of IRT available to those who might otherwise be using it in inappropriate situations.

Methods

This study used simulated data, based on item parameters obtained from an archival file of actual examinee responses. The use of real test data to generate initial item parameters, rather than arbitrarily setting parameter values or distributions, is intended to provide more generalizable results than simulated data may be able to yield. A 120-item certification exam was selected as the basis for the simulation. Although it was desirable to investigate the effect of a lengthy test, it was not deemed necessary to use all 120 items from this exam. Additionally, some computer programs intended for use in this study had limitations on the maximum number of items. Therefore a subset of 80 items was identified based on classical item indices (difficulty and discrimination indices), IRT discrimination indices, and factor loadings. This resulted in a set of 80 essentially unidimensional items. (Stout, 1987).

To model the complexity and variability of real data, a multidimensional approach was taken. In general, even when the assumption of essential unidimensionality is met, strict unidimensionality is not present in the observed data. In fact, as Davey, Nering, & Thompson, (1997) point out, the dominant factor found in many analyses of well constructed cognitive tests

typically accounts for less than half the score variance. Fitting a multidimensional latent trait model to observed data provides for a more accurate simulation of item response data by allowing a richer measure of the multiple skills and abilities which examinees often utilize in responding to an assessment task. When this approach is used in simulations the fitted model is not interpreted directly, but is "simply treated as a template from which new data can be generated" (Davey, Nering, & Thompson, 1997, p. 6). The purpose of using a multidimensional model to generate simulated data is thus simply to obtain the best reflection of real data possible.

The simulated data were generated by first fitting a 6-dimensional model to the scored responses of 2,862 examinees to the 80 items culled from the archival certification exam, using the multidimensional calibration program NOHARM (Fraser & McDonald, 1988). (NOHARM does not estimate *c* parameters, so these values were fixed in NOHARM at estimates obtained by first fitting a unidimensional model in BILOG.) Six dimensions were modeled in order to use the maximum number of dimensions which this version of NOHARM could fit. This yielded a set of item parameter estimates for each item, consisting of six discrimination parameters, one difficulty parameter, and one lower asymptote. Multiple examinee attributes were modeled through a set of six true thetas generated for each examinee from independent normal ability distributions (i.e., $N(0,1)$). These multiple examinee abilities were not interpreted directly, nor was one of greater importance than another. Rather, they were merely intended to model the multiple abilities which examinees typically bring to an assessment situation.

These ability and item parameters were regarded as true parameters for purposes of the study. They were used to generate item response vectors by determining the probability of a correct response on a given item, for a given examinee (based upon the set of item and theta parameters), and then comparing that probability to a random number sampled from a uniform (0,1) distribution. If the random number was less than or equal to the probability of a correct response, then the response was scored as correct.

The multidimensional approach to simulating data, and the degree of correspondence between the simulated data and the original archival data were next evaluated. A dataset consisting of the item response vectors for 2,862 simulees (i.e., a number equivalent to that in the archival dataset) was generated according to the procedures just described. Another simulated dataset of 2,862 simulees was also generated, using unidimensional item parameters obtained

from the archival data. Classical item difficulty and point-biserial discrimination indices were then computed on these three datasets (the archival data, the multidimensionally simulated data, and the unidimensionally simulated data). Plots of these values (Figures 1 and 2) reveal a much closer correspondence between the real, archival data and the multidimensionally simulated data than is found between the archival data and the unidimensionally simulated data. The improved accuracy of the data simulated under the multidimensional approach lends support for its use in this investigation.

Insert Figures 1 and 2 about here

One hundred datasets were then generated for each of the four sample sizes under investigation. The generated datasets were used as independent samples from a representative population of interest, and ability and item parameters for each of the studied models were estimated. These parameter estimates were obtained from the generated response vectors through the calibration program BILOG (Mislevy & Bock, 1990).

Models under investigation included the typical 1-parameter, 2-parameter, and 3-parameter models as benchmarks. Three additional, modified models consisted of a 2-parameter model with a restricted *a* parameter (i.e., a strong prior distribution was imposed), a 3-parameter model with a restricted *a* parameter, and a 3-parameter model with both a restricted *a* parameter and a common *c* parameter. This yielded a total of six models, three of which were unrestricted, and three of which were restricted.

The benchmark 1-PL model constrained all *a* parameters to be equal; both the 1-PL and the 2-PL models set the *c* parameters to zero (i.e., did not estimate the *c* parameters). The benchmark 2-PL and 3-PL models used BILOG's default prior distribution for *a* parameters, which is $.5^2$ in the lognormal metric (or, μ_a=1.13 and σ_a=.36 in the *a* metric). This default prior is typically imposed to avoid the extreme values sometimes estimated for *a* parameters (i.e., to prevent Heywood cases). For the benchmark 3-PL model, the default beta prior was also used for estimation of the *c* parameter. All three modified models imposed more informative priors on the *a* parameters. These modified 2-PL and 3-PL models included a prior of $.25^2$ in the lognormal metric (or, μ_a=1.03 and σ_a=.07 in the *a* metric). One modified 3-PL model also

constrained the *c* parameters to be equal to one another (but free to be non-zero). These modified models are noted as 2-PL*a*, 3-PL*a*, and 3-PL*ac*.

Each of these models was investigated with sample sizes of 1000, 500, 250, and 100. The largest sample size here is typically considered adequate for the 3-parameter model, while the smallest sample size might prove challenging for even the 1-parameter model. The full study was a 6 x 4 design, with the six models and four sample sizes yielding a total of 24 conditions. One hundred samples of each size were generated, and the results were analyzed across replications, to control for sampling error. Use of the same set of 100 samples across models (for a given sample size), provides the opportunity to make direct comparisons of model performance.

After the initial analysis of the simulated data, some samples failed to converge (although BILOG's options for number of EM cycles and Newton-Gauss iterations had been set to the values of 50 and 10 respectively). This was particularly true of sample size 100, where the only conditions with less than 5% nonconvergence were the 1-PL model and the 2-PL*a*. For sample size 100, the unmodified 2-PL model had 47 samples out of 100 not converge, the unmodified 3-PL model had 44 samples not converge, the 3PL*a* had 21 samples not converge and the 3-PL*ac* had 9 samples not converge. A decision was made to drop the 2-PL, 3-PL, 3-PL*a* and 3PL*ac* models for this smallest sample size based on the high rate of nonconvergence and the severe limitations a sample size of 100 imposes on the estimation algorithms used in models incorporating multiple parameters. For sample size 100, all samples converged for the 1-PL model and only two samples failed to converge for the 2PL*a* model. For those models and sample sizes remaining in the study which had nonconverging samples, new simulated data were generated from the original parameters and were used to replace these nonconverging samples in the raw data files.

Analyses

A variety of evaluative measures are conducted in analyses of item parameter recovery. The relative success of the six IRT models in this study was determined using indices of model-data fit, indices of the stability of the models, and indices of the relative accuracy of the models. In addition to measures of fit, stability, and accuracy of the models obtained from the calibration

samples, a crossvalidation of the models was conducted. Such a crossvalidation provides evidence of the accuracy and fit of the parameter estimates obtained in one sample, when they are applied to new data.

Fit Indices

Two indices of fit were calculated for each item in each sample. These fit indices represent the extent to which each model was able to predict the observed data in the sample. First, raw residuals from ability groups (Hambleton & Swaminathan, 1985) were calculated for each sample and each model. In this grouped residuals method, the range of estimates of theta in the sample is divided into ten equal intervals. Within each interval, the squared difference between the actual proportion of examinees who answered the item correctly and the expected proportion based on the IRT model of the item is calculated. The sum of these squared residuals, across the ten intervals, is calculated as the index of fit for the item in the sample, and the mean of these fit indices across the 100 samples was used:

$$r_i = \frac{\sum_k \left[\sum_j \left(\hat{\pi}_{ijk} - E_{ijk} \right)^2 \right]}{100}$$

where

r_i = raw residual for item i

$\hat{\pi}_{ijk}$ = observed proportion of correct responses for item i, interval j, and sample k,

E_{ijk} = predicted proportion of correct responses for item i, interval j, and sample k.

For the second index of fit, individual person residuals were calculated. This index provides a measure of the *average person fit*, while the first, grouped residuals method provides a measure of the *fit to average people*. The person residuals are the residuals between the observed item data and the obtained probabilities (X_{ij} - P_{ij}) calculated for each item and each examinee[1]. The average of these residuals across examinees is used as the fit index for the item:

[1] Our thanks to Tim Davey of ACT for suggesting this approach.

$$XP = \frac{\sum_{k}\left[\frac{1}{n_j}\sum_{j}(X_{ijk} - P_{ijk})^2\right]}{100}$$

where

XP_i = mean person residual for item i

X_{ijk} = observed response for item i, examinee j, and sample k,

P_{ijk} = estimated probability of correct responses for item i, examinee j, and sample k,

n_j = number of examinees in the sample.

Accuracy Indices

Because the data were generated from a 6-dimensional model, the accuracies of the item parameter estimates obtained from unidimensional models cannot be obtained (that is, the population from which the samples were generated does not have item parameters corresponding to the *a* and *b* parameter estimates obtained in the analyses of the samples). Similarly, the population is not characterized by a single theta value to which sample estimates of theta may be compared. However, the characteristics of the population (the known item parameters from the 6-dimensional model and the known theta values for each simulee on each dimension) provide a known probability of correct response to each item for each examinee. These known probabilities represent the "truth" from which the relative accuracies of the sample estimates may be obtained.

Three indices of the accuracy of the IRT models were used. First, the accuracy of the individual probabilities of correct response to each item for each examinee were compared to the known, true probabilities obtained from the population parameters. That is, the degree to which the *estimated* response probabilities reflected the *expected* response probabilities was taken as a measure of a model's accuracy. The mean squared error (MSE) of these probability estimates was used as an index of accuracy at the item level. This statistic is given as

$$MSE_i = \frac{\sum_k \left[\frac{1}{n_j} \sum_j \left(T_{ijk} - P_{ijk} \right)^2 \right]}{100}$$

where

MSE_i = mean squared error for item i,

T_{ijk} = true probability of correct response for item i, examinee j, and sample k,

P_{ijk} = estimated probability of correct responses for item i, examinee j, and sample k.

A second index of accuracy was obtained at the person level. For this index, the sample estimated number correct score for each examinee (the sum, over items, of the estimated probabilities of correct response) was compared to the corresponding true score from the population (the sum, over items, of the true probabilities of correct response). The root mean squared error (RMSE) of this value, over examinees in each sample, was used as the index

$$RMSE_{nc} = \frac{\sum_k \left[\frac{1}{n_j} \sum_j \left(\sum_i T_{ijk} - \sum_i P_{ijk} \right)^2 \right]^{\frac{1}{2}}}{100}$$

where

$RMSE_{nc}$ = root mean squared error, number correct

T_{ijk} = true probability of correct response for item i, examinee j, and sample k,

P_{ijk} = estimated probability of correct responses for item i, examinee j, and sample k.

The final index of accuracy was simply the Spearman rank correlation between the sample estimated number correct score and the known population value of the number correct score.

Stability Indices

Estimates of the stability of the item parameter estimates and the item response functions were obtained by calculating the standard deviations of the estimates of the *a* and *b* parameters, and the standard deviations of the entire item characteristic curve (ICC) over the 100 samples. The standard deviations of the item parameter estimates were obtained using the usual formula for the sample estimate of a population standard deviation:

$$\sigma_{ij} = \sqrt{\frac{\sum_{k} \left(X_{ijk} - \mu_{ij} \right)^2}{99}}$$

where

σ_{ij} = standard deviation of parameter i, for item j,

X_{ijk} = estimate of parameter i, for item j, in sample k, and

μ_{ij} = mean of parameter i, for item j in the 100 samples.

The standard deviation of the entire ICC was obtained by dividing the theta scale into 31 equally spaced intervals (spanning a theta range from -3.0 to 3.0) and calculating the expected proportion of correct responses within each interval (P_{mn}, for interval m and item n), given the item parameter estimates obtained from the sample data. The standard deviation was then obtained as

$$\sigma_n = \frac{\sum_{m} \sqrt{\frac{\sum_{o} \left(P_{mno} - \mu_{mn} \right)^2}{99}}}{31}$$

where

σ_n = standard deviation of item characteristic curve for item n,

P_{mno} = estimate of proportion of correct responses for interval m, item n, and sample o,

μ_{mn} = mean of estimates for interval m, item n, in the 100 samples.

Crossvalidation Approach

In addition to examining the IRT models in the samples for which model estimates were obtained, the fit and accuracy of the models were evaluated in crossvalidation samples. The procedure of crossvalidation has been used extensively in statistical contexts such as regression and discriminant function analyses (see, for example, Mosier, 1951; Camstra & Boomsma, 1992; Mosteller & Tukey, 1977; Picard & Cook, 1984). The rationale underlying the need for crossvalidation is that parameter estimates (e.g., in regression or discriminant function analyses) provide a model that fits the sample from which the estimates were derived (the calibration sample) better than the model will fit in either new samples or in the entire population from which the calibration sample was obtained. Idiosyncratic aspects of specific samples are capitalized upon when models are estimated and the resulting indices of fit (for example, R^2 or canonical correlation coefficients) are overly "optimistic." Although crossvalidation of model parameter estimates has typically not been investigated in IRT methodological research, the applications of IRT models in real-world measurement involves the use of item parameter estimates obtained in one sample (when items are calibrated) to new samples (when calibrated item banks are used in applied testing programs). Crossvalidation of the IRT models investigated in this study was conducted by generating 20 new samples (each with 100 examinees) and applying each set of sample item parameter estimates to each of these new samples.

Results

Model-Data Fit in Calibration Samples

The two fit indices obtained from each model are presented in Table 1. Reported in the table are average fit indices across the 80 items. The standard deviations in the table are the average standard deviation in item fit across the 100 samples. The grouped residuals fit indices are graphed in Figures 3-5, while the person residuals are graphed in Figures 6-8, for the models incorporating 1 parameter, 2 parameters, and 3 parameters, respectively. (These and following figures do not show the four models which were dropped from the analysis for the sample size of 100, due to high numbers of nonconverging replications. The figures display results for both

calibration and crossvalidation data; the crossvalidation results will be discussed later in the paper.)

Insert Table 1 and Figures 3-8 about here

The tabled results for grouped residuals indicate that the unmodified 2-PL evidenced the smallest residuals of the six models examined, while the 1-PL displays the largest residuals (or, the poorest fit). The modified 2-PL*a* shows slightly poorer fit than the unmodified 2-PL, while the two modified 3 parameter models (3-PL*a* and 3-PL*ac*) display better fit than the unmodified 3-PL. Marked improvements in fit can be noted as sample size increases, with the greatest improvement appearing for all sample sizes above 100. With the exception of the 1-PL model, the models display some convergence at the largest sample size.

For the individual person residuals, across all sample sizes the models including 3 parameters demonstrate better fit than those including only 2 parameters, while the 1-parameter model again displays the poorest fit. Best fit is provided by the unconstrained 3-PL model, followed closely by the 3-parameter model with constrained *a* (3-PL*a*). The unmodified 2-PL again shows slightly better fit than the modified 2-PL*a*.

The variation in the fit of the models across the samples is reported in Table 1. The standard deviations reported in this table are the average standard deviations of the fit indices in the 100 samples. Small values of this statistic reflect consistency in fit across the samples, while large values reflect greater amounts of variation in fit with different samples of examinees. The standard deviations for the grouped residuals decreased as sample size increased, while values for the person residuals, on the other hand, remained very similar across sample size.

As the figures indicate, the two measures of fit display differing patterns across sample size. The grouped residuals method displays a decrease in values (or, an apparent improvement in fit) as sample size increased. This tendency might be explained by the improved estimate of proportion correct within a given range of estimated theta that is provided by the larger samples. Because these proportions are obtained as the mean item score (0,1) within the range, as sample size increases, the mean becomes a better estimate of this value. With small samples, some of these means are based on very few examinee responses. This apparent improvement in fit for the

grouped residuals, therefore, results not from better parameter estimates, but from better estimates of observed proportions. The individual person residual is based on a single item response (correct or not), regardless of the sample size. Therefore, no artifactual improvement in fit is evident at the person level. In fact, the consistency of this statistic across sample size suggests that it is a better measure of fit than the more typical grouped residuals approach.

Accuracy in Calibration Samples

The next general method for evaluating the success of the six models was the accuracy of sample estimates obtained. Estimates of accuracy were obtained by: (1) computing the mean squared error (MSE) between the estimated and expected response probabilities, (2) calculating the RMSE of the estimated and expected number correct scores, and (3) by computing Spearman rank correlation between the estimated and expected number correct scores. While the MSE is a measure of accuracy at the item level, the other two indices represent accuracy at the person level. These accuracy estimates are reported in Table 2 and are graphed for the models with 1 parameter, 2 parameters, and 3 parameters as: MSEs in Figures 9-11, RMSEs Figures 12-14, and rank correlations in Figures 15-17.

Insert Table 2 and Figures 9-17 about here

Considering only the results for the calibration data, accuracy of the models as measured by the MSE of the estimated and expected response probabilities displays a minimal effect for sample size (Figures 9-11). The model displaying the largest mean squared difference between the estimated and expected response probabilities is the 1-PL. The best performance, or the smallest MSE, is provided by the 3-PL and the 3-PL*a* models (overlaid in the figure), followed closely by the 3-PL*ac*. The unmodified 2-PL model displayed slightly less error than the 2-PL*a*, across sample size.

An examination of the RMSE of the estimated and expected number correct scores as a measure of accuracy also indicates little effect for sample size. All six models perform very similarly across the sample sizes considered in this study, and yield a similar pattern of results to that found under the MSE analysis. The most accurate models (e.g., those displaying the smallest RMSE) are the 3-PL and the 3-PL*a* models, followed closely by the 3-PL*ac*. Once

again, the unmodified 2-PL model displayed slightly less error than the 2-PL*a*, across sample size, while the poorest accuracy can be noted for the 1-PL model.

The final measure of accuracy was the Spearman rank correlations between the estimated and expected number correct scores. An examination of these results indicates that the 1-PL model yielded estimated number correct scores which correlated with true values markedly less well than those obtained from other models. The remaining models performed very similarly to one another, with the unmodified 3-PL model displaying slightly lower correlations.

Stability Across Calibration Samples

The final general method for evaluating the success of the six models for calibration data was the stability of item parameter estimates across samples. Indices of such stability were obtained by calculating the standard deviations of the estimates for the *a* and *b* parameters for each item, then averaging these standard deviations across the 80 items on the test. In addition, an overall measure of the stability of the item curves was obtained by calculating the standard deviation of P_{ij} at each of 31 theta values. These stability estimates are presented in Table 3.

Insert Table 3 about here

The reader is cautioned to avoid comparisons of stability across IRT models with different numbers of parameters. The indeterminacy of scale for the parameter estimates prohibits a direct comparison between, for example, the parameter estimates in the 1-PL and 2-PL models. However, comparisons between modified and unmodified models with the same number of parameters, and trends in the stability across sample sizes within models, are directly interpretable.

An examination of the stability of the estimates of the *b* parameter suggests that all of the estimates became more stable as sample size increased, and that the modified models were more stable than the unmodified counterparts (e.g., the *b*s estimated from the 2-PL*a* model were more stable than those estimated from the 2-PL model). Further, for the 3-PL models, the stability increased as more constraints were placed on the model (that is, the 3-PL*ac* model provided more stable *b* estimates than those of the 3-PL*a* model, which were more stable than those of the 3-PL model).

The stability of the estimates of the *a* parameter showed a somewhat different pattern from that obtained with the *b* parameter. Although an increase in stability was seen with increasing sample size, the degree of increase was less striking than that observed with the *b* parameter estimates. As with the *b* parameter estimates, the estimates of *a* obtained in the modified models were more stable than those obtained in their unmodified counterparts, differences which were maintained even with sample sizes as large as 1000. In contrast to the results observed with the estimates of *b*, however, the estimates of *a* obtained from the 3-PL*ac* model were more stable only with samples of size 1000. For the smaller samples observed, the *a* estimates from the 3-PL*a* model were more stable. However, either of these modified models provided consistently more stable estimates of *a* than the unmodified 3-PL model.

Finally, the estimates of the stability of the item characteristic curves were consistent with the previous stability indices. Across all of the models, stability increased with larger samples. As with the stability of the *b* estimates, the modified models were more stable than their unmodified counterparts, and for the 3-PL models, the more constrained modified model (3-PL*ac*) was more stable than the less constrained modified model (3-PL*a*).

Model-Data Fit in Crossvalidation Samples

Results for the model-data fit analyses on the crossvalidation data are reported in Table 4 and are displayed, along with the calibration data results, in Figures 3-8. The general tendency for poorer fit under crossvalidation is an unsurprising, but important finding. Overall, the 1-PL model displays good performance on crossvalidation data, for both the grouped residuals and person residuals indices of fit (Figures 3 and 6).

Insert Table 4 about here

An interesting finding of these analysis is that while the less constrained models fit better in the calibration samples, the more constrained models fit better under crossvalidation. For the models with 2 and 3 parameters, inclusion of constraints improves fit (e.g., the 2-PL*a* displays better fit than the unmodified 2-PL) as indicated by both measures of fit. This might suggest that the freedom to vary, which more parameters and fewer constraints provide, results in some capitalizing on chance, or overfitting of the model to the data.

Finally, for the 2 and 3 parameter models, the plots also indicate a small but noticeable pattern across sample size. For both indices, poorer fit is displayed at the largest sample size, suggesting a greater tendency towards fitting idiosyncrasies in the sample at these larger sizes.

Accuracy in Crossvalidation Samples

Accuracy was examined on the crossvalidation through the same three indices computed on calibration data: the MSE between the estimated and expected response probabilities, the RMSE of the estimated and expected number correct scores, and the Spearman rank correlations between the estimated and expected number correct scores. These results are reported in Table 5 and in Figures 9-17. Several findings under the fit analyses are supported by the results of the accuracy analyses.

Insert Table 5 about here

A direct comparison of accuracy as computed under crossvalidation and calibration data indicates a general tendency for lesser accuracy under crossvalidation conditions, for the models incorporating 2 and 3 parameters. An examination of the results for the MSE of the estimated and expected response probabilities, reveals that larger values, indicating decreased accuracy, are found for the crossvalidation data as compared to the calibration data, for the models with 2 and 3 parameters (Figures 10 and 11). The RMSE of the estimated and expected number correct scores, when calculated for the models with 2 and 3 parameters (Figures 13 and 14) displays this same pattern at the largest sample sizes, while better accuracy is found with the calibration data at the smallest sample size. Results for the Spearman rank correlations between the estimated and expected number correct scores also indicate lower correlation under crossvalidation conditions, for the 2 and 3 parameter models. The 1-PL model, as opposed to the other five models in the study, demonstrates improved performance on all three indices of accuracy, for crossvalidation data as compared to the original calibration data (Figures 9, 12, and 15).

Another pattern, similar to results seen with the fit indices, is found for the models with 2 and 3 parameters. Results for all three accuracy indices indicate that although the less constrained models fit better in the calibration samples, the more constrained models fit better under crossvalidation. This may be an important finding, with implications for practical

applications where item parameter estimated are obtained from one sample, and then used repeatedly on other samples.

Finally, the same pattern across sample size noted for the fit analyses is found for the 2 and 3 parameter models under the accuracy analyses. That is, decreased performance is found at the largest sample size.

Discussion

This study was designed to investigate the relative effects of sample size and model selection on item parameter estimation, and whether various modifications to typical models might improve estimation under small sample conditions. The three modified models, along with the three unmodified models, were compared in terms of fit, stability, and accuracy. The same six models had been investigated in an earlier study (Parshall, Kromrey, and Chason, 1996), which used a moderate length (i.e., 40-items) achievement exam and a unidimensional data generation technique.

Calibration Data

One strong pattern of results for calibration samples in this study (as well as the earlier research) was the tendency for models which displayed the best fit within samples, to display the poorest stability across samples. Conversely, models which demonstrated good stability across replications tended to be associated with relatively poorer fit within replications. Such a result should be anticipated. The IRT models with fewer constraints are free to establish parameters which best fit the calibration sample, leading to better fit than that obtained with more constrained models. However, this freedom to better fit the parameters to the sample results in more variability in the parameters across samples, resulting in lower levels of stability than was evidenced with the more constrained models.

The addition of constraints to the models tended to improve stability, while decreasing both fit and accuracy, in comparison to the unconstrained models with the same number of parameters (e.g., the 2-PL*a* as compared to the 2-PL). In the earlier study of a moderate-length test, imposing a more informative prior on the variance of the *a* parameter tended to improve both fit and stability. In Harwell and Janosky's (1991) investigation into the effect of differing prior variances on the *a* parameter in a 2-PL model, more informative priors were found to

improve parameter recovery with small samples and short tests. (Results were evaluated by computing the root mean square difference between true and estimated item parameters.)

The addition of constraints also aided the estimation process sufficiently that when datasets were calibrated under modified models, BILOG's convergence criteria were more likely to be met. For example, when the same set of 100 samples of size 100 were calibrated with the 3-PL model, 56 samples converged. With the addition of a constraint on the a parameter (the 3-PL*a*), 79 converged. And, for the 3-PL*ac* model, 91 samples converged. While all of these models were dropped from further analysis due to the relatively high proportion of nonconvergence, this pattern of effect due to model modification was also evident at the larger sample sizes.

For the smallest sample size investigated here (100), only the 1-PL and 2-PL*a* were retained for full analysis. For these two models at this sample size, the 1-PL yielded poorer fit, better stability, and nearly identical accuracy as compared to the 2-PL*a*.

Another interesting finding in this study was that the unconstrained 3-PL yielded the most accurate parameter estimates for all sample sizes of 250 and above. In the earlier study of these modified models, the data were generated according to the 3-PL model, leading to the possibility that model comparisons would be biased in favor of models incorporating 3-parameters. In this study, data were generated by a more complex, multidimensional approach. This approach should yield realistic data, without bias towards any of the models investigated. Both studies found positive results for fit for the models incorporating 3-parameters; this study additionally found the set of 3-parameter models to yield the most accurate results. While it may not be surprising that the more flexible models are better fitting, it is surprising that the better fit is translated into more accurate parameters, even at sample sizes as small as 250.

Crossvalidation Data

Crossvalidation is a methodological approach which has not often been used in studies of IRT parameter recovery. Nevertheless, it is a highly appropriate technique, due to the correspondence between the analysis of parameter estimate performance in new data and the way in which calibration data are often used in practice. In this study, crossvalidation data were used to investigate fit and accuracy of the item parameter estimates when used with new data.

A comparison of results for calibration data to crossvalidation data in some ways parallels the comparison of results for the calibration data measures of fit and stability. Models which display better fit to a given sample tend to display poorer stability across samples. In a similar manner, models which display better performance on calibration data, for both fit and accuracy, tend to display poorer performance on crossvalidation data. The 1-PL model, which showed generally poor fit and accuracy under many of the calibration conditions, displayed improved performance under the crossvalidation conditions. Constraints added to the 2-PL and 3-PL models, which in some cases degraded fit and accuracy for calibration data, definitely improved both fit and accuracy for the crossvalidation data.

Summary

These results suggest that further investigation into model modifications are worthwhile. Practitioners especially will be benefited if suitable modifications are found which enable the estimation of better fitting, more stable, and more accurate item parameters under limited sample size conditions.

Additionally, several methodological approaches included in this study appear worthy of further investigation. These approaches are: the use of MIRT data simulation techniques, the individual person fit index, and crossvalidation methods for evaluating parameter recovery.

References

Barnes, L. B., & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4*, 143-157.

Camstra, A. & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: An overview. *Sociological Methods and Research, 21*, 89-115.

Davey, T., Nering, M., & Thompson, T. (1997, March). *Realistic simulation of item response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

de Jong, J. H. A. L., & Stoyanova, F. (1994, March). *Theory building: Sample size and data-model fit.* Paper presented at the annual Language Testing Research Colloquium.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7*, 171-186.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*, 143-155.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Hulin, Lissak, & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. & Bock, R. D. (1990). *Bilog 3: Item analysis and test scoring with binary logistic models*. Chicago: Scientific Software.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11*, 5-11.

Mosteller, F. & Tukey, W. J. (1968). Data analysis, including statistics. In G. Lindzy, E. Arnun (Eds.), *Handbook of Social Psychology (2nd ed.)*. Reading, MA: Adison-Wesley.

Parshall, C. G., Kromrey, J. D., & Chason, W. (1996, June). *Comparison of alternative models for item parameter estimation with small samples*. Paper presented at the annual meeting of the Psychometric Society, Banff, Canada.

Patsula, L. N., & Pashley, P. J. (1996, April). *Pretest item analyses using polynomial logistic regression: An approach to small sample calibration problems associated with computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Picard, R. R. & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79*, 575-583.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Sireci, S. G. (1992, August). *The utility of IRT in small-sample testing applications*. Paper presented at the annual meeting of the American Psychological Association, Washington, D.C.

Spray, J., Kalohn, J. C., Schulz, M., & Fleer, P., Jr. (1995, June). *The effect of model misspecification on classification decisions made using a computerized test: 3-PL versus 1-PL logistic item response models*. Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Stone, C. A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4*, 125-141.

Stout, William (1987). A statistical approach for determining the latent trait dimensionality in psychological testing. *Psychometrika, 52*, 589-617.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item repsonse models. In P.W. Holland & H. Wainer (Eds.) *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press. .

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Table 1

Indices of Model-Data Fit for Six Models and Four Sample Sizes, for Calibration Data.

		Fit Index			
		Person Residual		Grouped Residual	
Sample Size	Model	Mean	SD	Mean	SD
100	1-PL	0.1512	0.2541	0.3378	0.2764
	2-PLa	0.1500	0.2509	0.3001	0.3055
250	1-PL	0.1539	0.2561	0.2238	0.1759
	2-PL	0.1506	0.2463	0.1435	0.1151
	2-PLa	0.1515	0.2503	0.1648	0.1360
	3-PL	0.1468	0.2225	0.1981	0.1849
	3-PLa	0.1471	0.2229	0.1921	0.1669
	3-PLac	0.1484	0.2322	0.1887	0.1800
500	1-PL	0.1544	0.2565	0.1866	0.1370
	2-PL	0.1511	0.2452	0.1016	0.0618
	2-PLa	0.1517	0.2484	0.1174	0.0781
	3-PL	0.1475	0.2227	0.1597	0.1538
	3-PLa	0.1476	0.2232	0.1317	0.1163
	3-PLac	0.1487	0.2313	0.1376	0.1386
1000	1-PL	0.1550	0.2569	0.1670	0.1221
	2-PL	0.1517	0.2446	0.0848	0.0592
	2-PLa	0.1520	0.2467	0.0927	0.0628
	3-PL	0.1483	0.2227	0.1105	0.1051
	3-PLa	0.1483	0.2233	0.0927	0.0787
	3-PLac	0.1492	0.2306	0.1077	0.1144

Table 2

Indices of Accuracy Estimates for Six Models and Four Sample Sizes, for Calibration Data.

Sample Size	Model	Accuracy				
		RMSE		MSE		Spearman Correlation
		Mean	SD	Mean	SD	
100	1-PL	5.33024	0.33257	0.0235	0.0373	0.9447
	2-PLa	5.33431	0.35248	0.0236	0.0361	0.9470
250	1-PL	5.37663	0.20558	0.0225	0.0362	0.9467
	2-PL	5.08551	0.20902	0.0199	0.0325	0.9514
	2-PLa	5.26371	0.21038	0.0207	0.0334	0.9511
	3-PL	4.07723	0.20331	0.0166	0.0301	0.9505
	3-PLa	4.15186	0.20428	0.0166	0.0299	0.9510
	3-PLac	4.41505	0.22480	0.0174	0.0308	0.9512
500	1-PL	5.36537	0.13937	0.0221	0.0358	0.9474
	2-PL	4.99113	0.14974	0.0189	0.0313	0.9519
	2-PLa	5.14727	0.14710	0.0195	0.0320	0.9521
	3-PL	4.00629	0.12961	0.0159	0.0292	0.9514
	3-PLa	4.07678	0.12963	0.0159	0.0291	0.9519
	3-PLac	4.32072	0.14358	0.0167	0.0298	0.9521
1000	1-PL	5.37212	0.10270	0.0220	0.0358	0.9490
	2-PL	4.93328	0.10363	0.0183	0.0306	0.9536
	2-PLa	5.04709	0.10283	0.0187	0.0311	0.9539
	3-PL	3.96871	0.09254	0.0155	0.0286	0.9532
	3-PLa	4.02315	0.09581	0.0155	0.0286	0.9536
	3-PLac	4.24992	0.10212	0.0162	0.0293	0.9537

Table 3

Indices of Stability of Estimates for Six Models and Four Sample Sizes, for Calibration Data.

Sample Size	Model	Stability		
		b	a	ICC
100	1-PL	0.3887	0.0751	0.0443
	2-PLa	0.5625	0.1203	0.0520
250	1-PL	0.2289	0.0425	0.0268
	2-PL	0.3016	0.2031	0.0421
	2-PLa	0.2409	0.1240	0.0342
	3-PL	0.3219	0.2415	0.0393
	3-PLa	0.2898	0.1182	0.0352
	3-PLac	0.2641	0.1280	0.0321
500	1-PL	0.1653	0.0295	0.0195
	2-PL	0.2275	0.1608	0.0323
	2-PLa	0.1895	0.1168	0.0277
	3-PL	0.2444	0.2122	0.0315
	3-PLa	0.2246	0.1222	0.0286
	3-PLac	0.2000	0.1269	0.0251
1000	1-PL	0.1147	0.0225	0.0137
	2-PL	0.1677	0.1232	0.0239
	2-PLa	0.1463	0.1010	0.0216
	3-PL	0.1918	0.1806	0.0252
	3-PLa	0.1772	0.1207	0.0231
	3-PLac	0.1512	0.1148	0.0194

Table 4

Indices of Model-Data Fit for Six Models and Four Sample Sizes, for Crossvalidation Data.

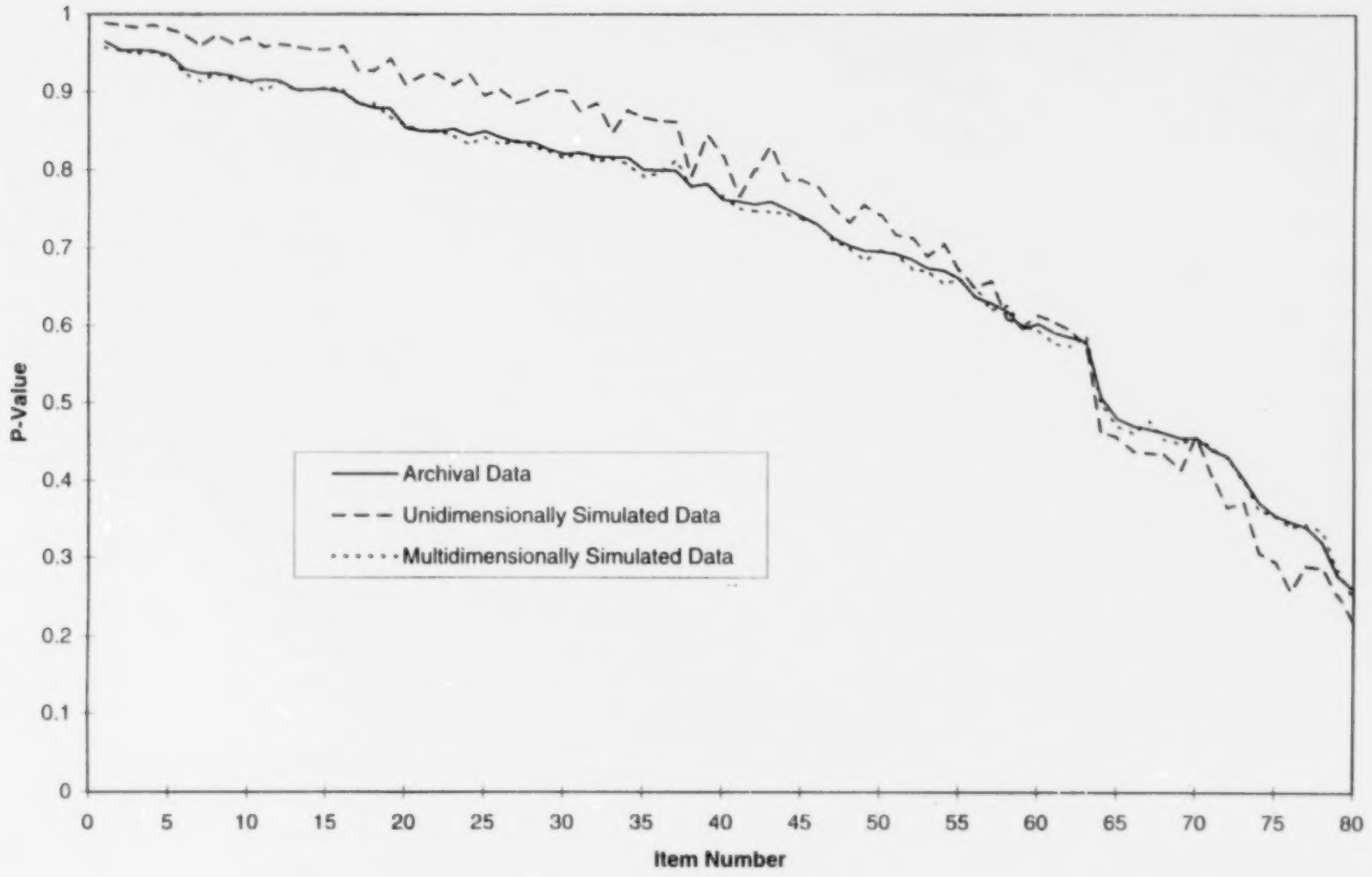
Sample Size	Model	Fit Index: Person Residual Mean	Fit Index: Person Residual SD	Fit Index: Grouped Residual Mean	Fit Index: Grouped Residual SD
100	1-PL	0.1574	.2548	0.4337	.3703
	2-PLa	0.1562	.2412	0.3323	.3048
250	1-PL	0.1540	.2412	0.3323	.2494
	2-PL	0.1556	.2404	0.4848	.4493
	2-PLa	0.1547	.2384	0.4143	.4005
	3-PL	0.1539	.2305	0.3965	.3730
	3-PLa	0.1533	.2257	0.3873	.3541
	3-PLac	0.1530	.2238	0.3775	.3400
500	1-PL	0.1536	.2412	0.3074	.2097
	2-PL	0.1571	.2440	0.3997	.3617
	2-PLa	0.1555	.2401	0.3525	.3258
	3-PL	0.1543	.2321	0.3445	.3083
	3-PLa	0.1536	.2273	0.3402	.2969
	3-PLac	0.1511	.2153	0.3201	.2455
1000	1-PL	0.1521	.2277	0.3113	.2311
	2-PL	0.1644	.2506	0.4699	.3808
	2-PLa	0.1620	.2467	0.4647	.4001
	3-PL	0.1596	.2383	0.4346	.3791
	3-PLa	0.1579	.2327	0.4115	.3618
	3-PLac	0.1568	.2297	0.3947	.3494

Table 5

Indices of Accuracy Estimates for Six Models and Four Sample Sizes, for Crossvalidation Data.

Sample Size	Model	Accuracy				
		RMSE		MSE		Spearman Correlation
		Mean	SD	Mean	SD	
100	1-PL	4.09	2.18	.0236	.0433	.9301
	2-PLa	3.39	0.66	.0024	.0346	.9442
250	1-PL	3.40	0.68	.0203	.0333	.9408
	2-PL	4.62	2.02	.0217	.0424	.9326
	2-PLa	4.19	1.77	.0208	.0385	.9373
	3-PL	4.12	1.56	.0200	.0366	.9381
	3-PLa	4.07	1.41	.0194	.0353	.9390
	3-PLac	4.04	1.29	.0191	.0345	.9398
500	1-PL	3.33	0.27	.0199	.0323	.9421
	2-PL	5.05	3.47	.0233	.0447	.9292
	2-PLa	4.49	2.94	.0217	.0397	.9354
	3-PL	4.36	2.57	.0205	.0374	.9365
	3-PLa	4.27	2.31	.0198	.0358	.9377
	3-PLac	3.87	0.36	.0173	.0301	.9446
1000	1-PL	3.60	0.42	.0184	.0310	.9433
	2-PL	7.44	5.86	.0310	.0622	.9034
	2-PLa	6.87	5.31	.0284	.0575	.9118
	3-PL	6.27	4.90	.0260	.0519	.9174
	3-PLa	5.86	4.57	.0243	.0480	.9217
	3-PLac	5.56	4.29	.0231	.0452	.9249

Figure 1 - Comparison of Classical Item Difficulty Indices for Real and Simulated Data



31

Figure 2 - Comparison of Classical Item Discrimination Indices for Real and Simulated Data

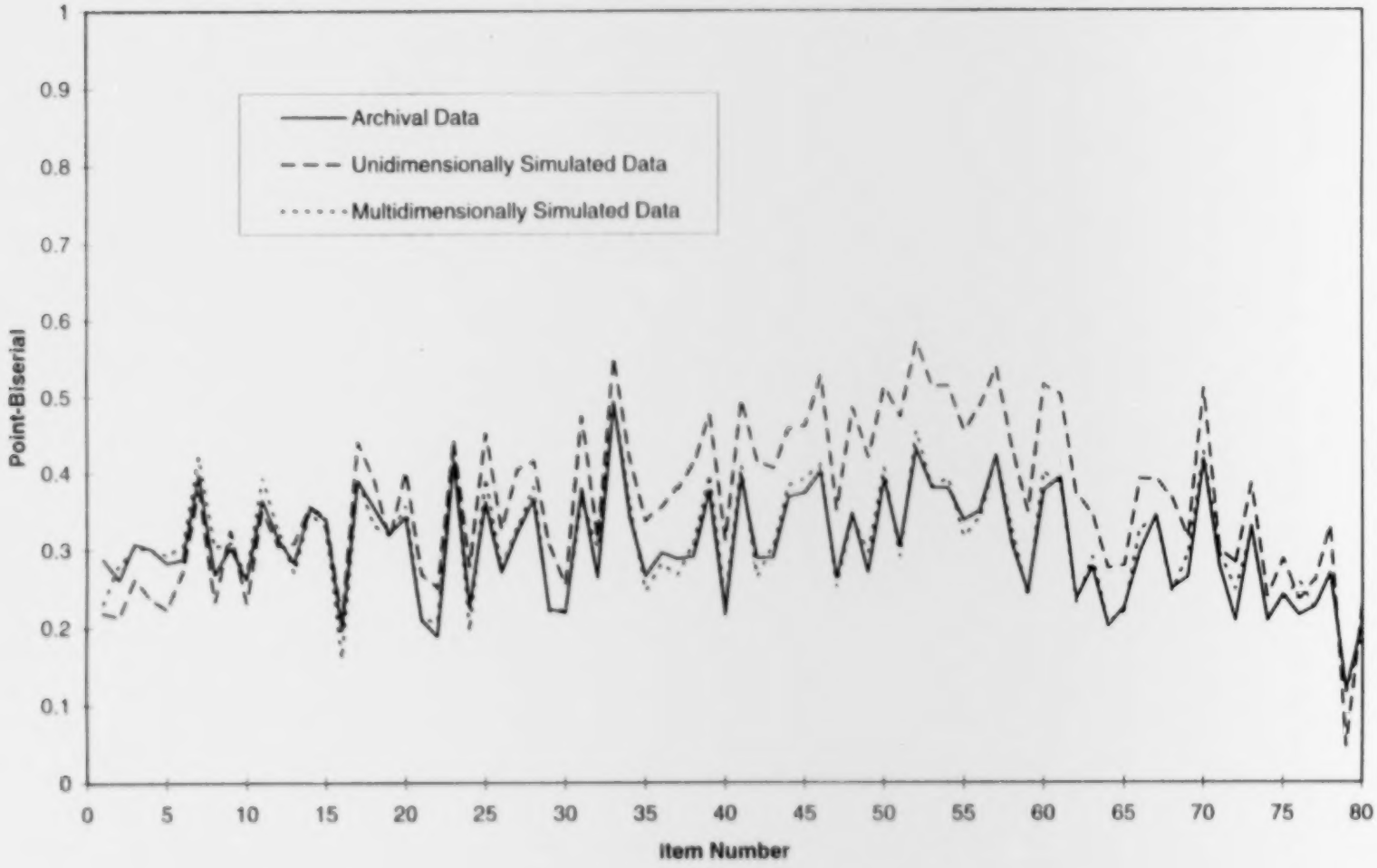


Figure 3 - Mean Grouped Residuals (1-PL model)

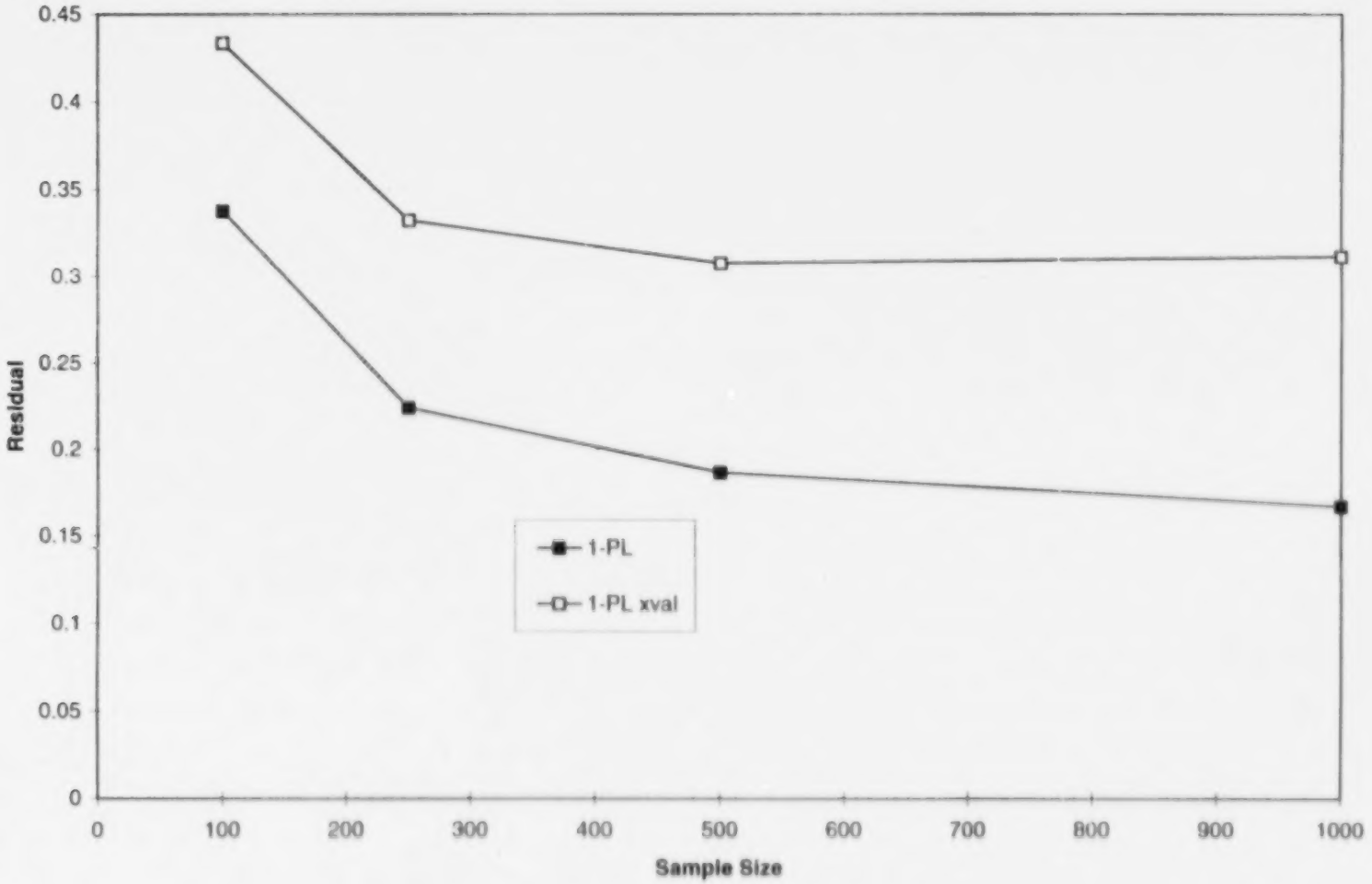


Figure 4 - Mean Grouped Residuals (2PL models)

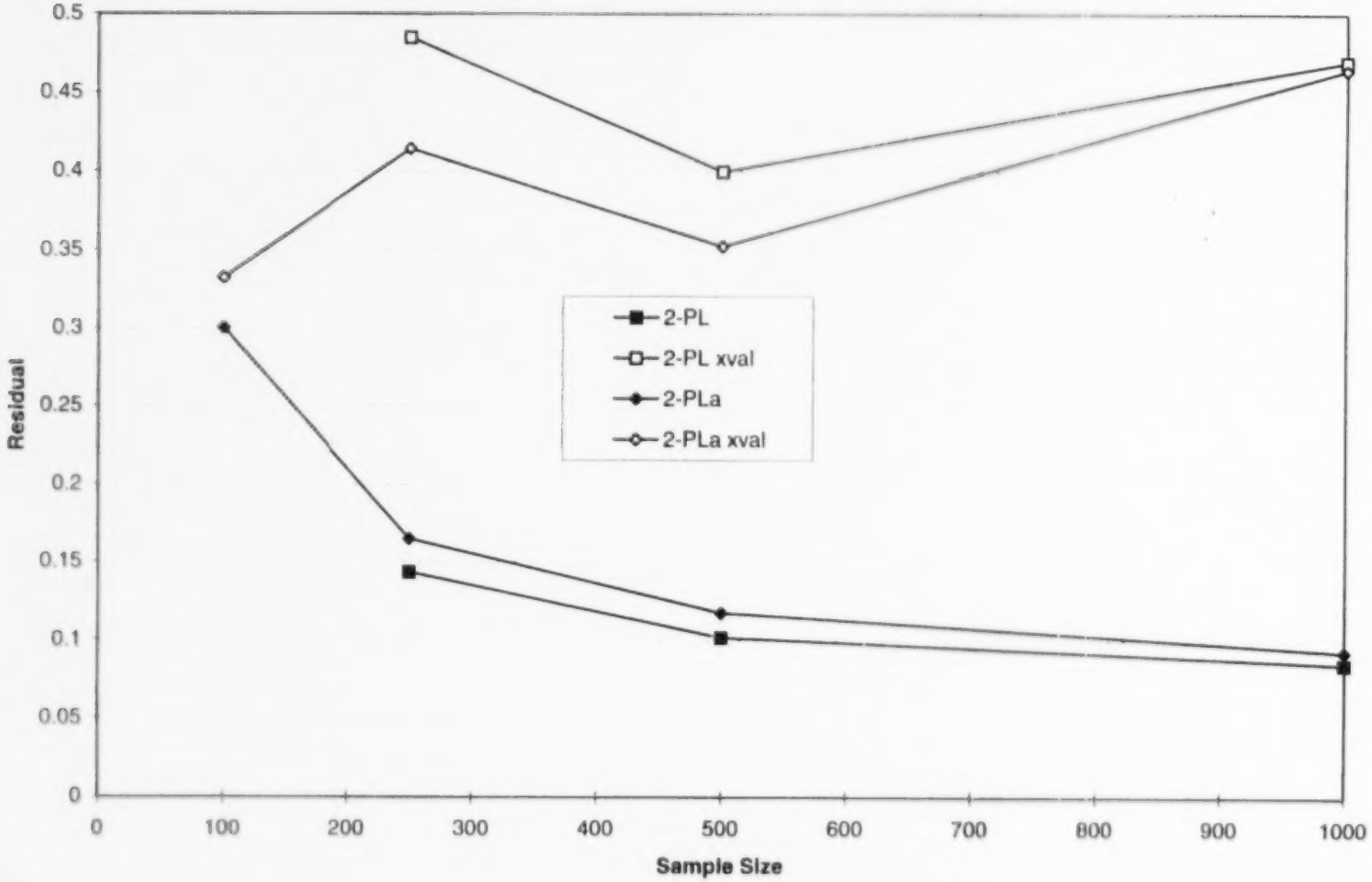


Figure 5 - Mean Grouped Residuals (3-PL models)

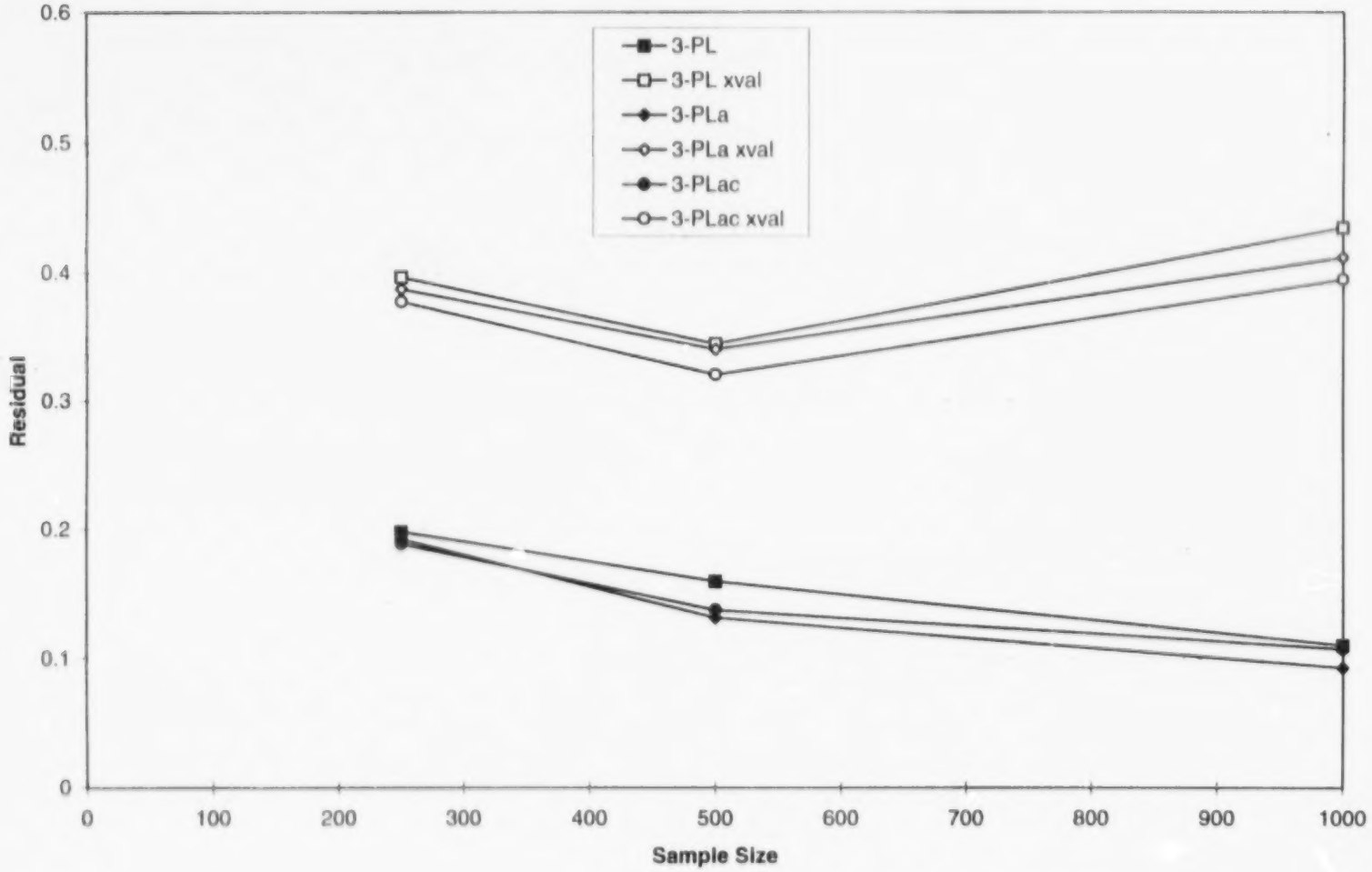


Figure 6 - Mean Person Residuals (1-PL model)

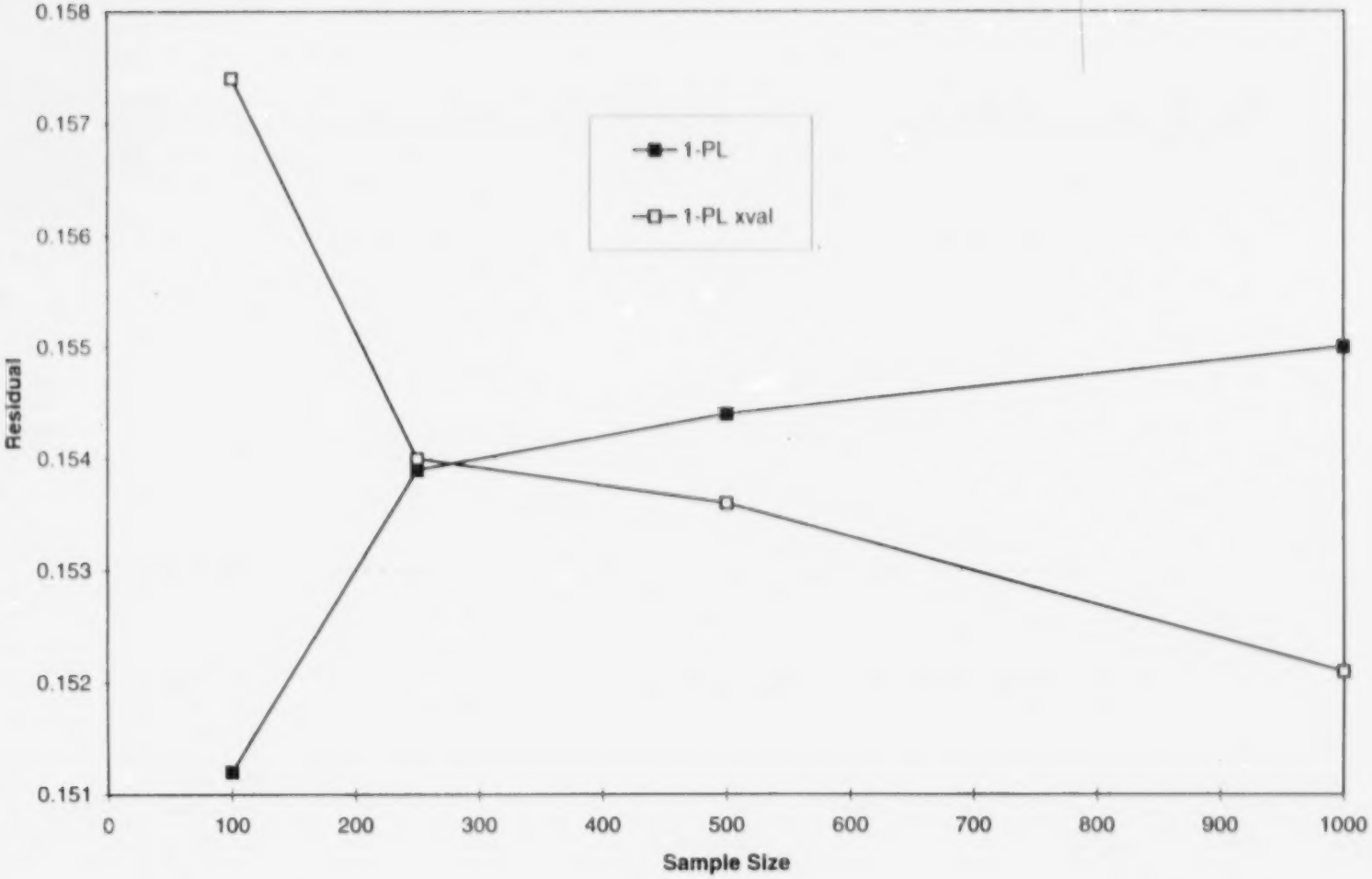


Figure 7 - Mean Person Residuals (2-PL models)

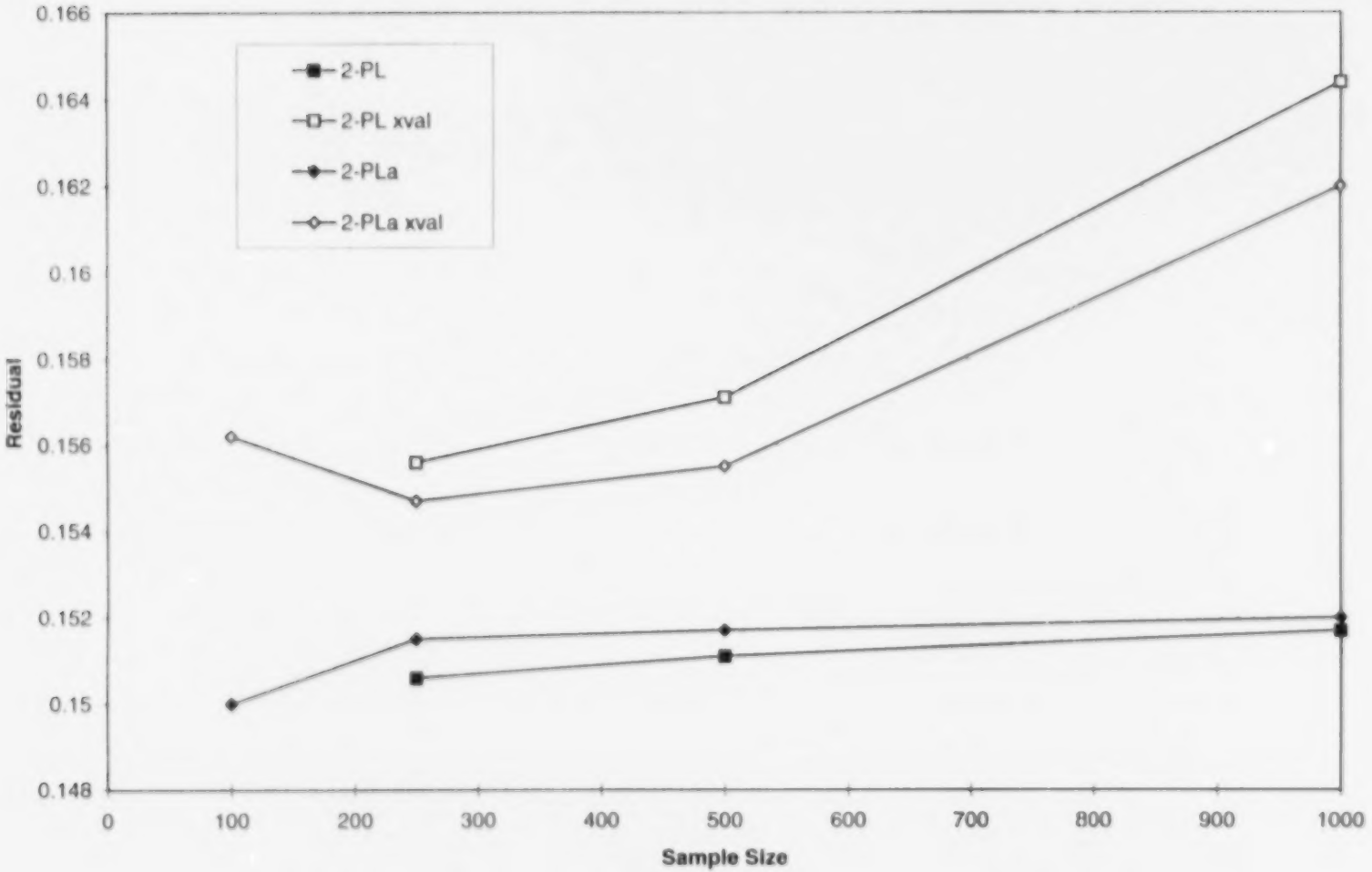


Figure 8 - Mean Person Residuals (3-PL models)

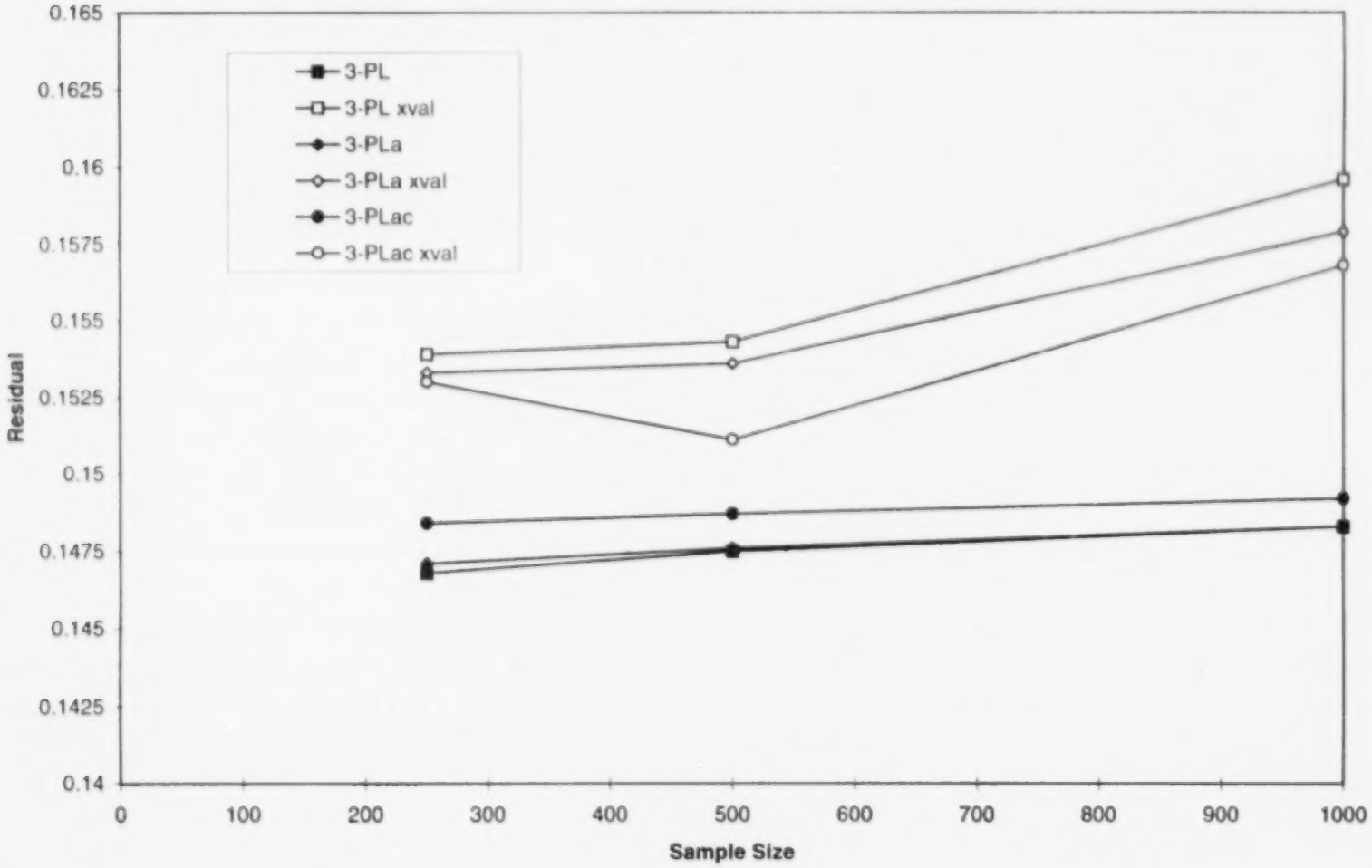


Figure 9 - Mean Squared Error Between Estimated and Expected Response Probabilities (1-PL model)

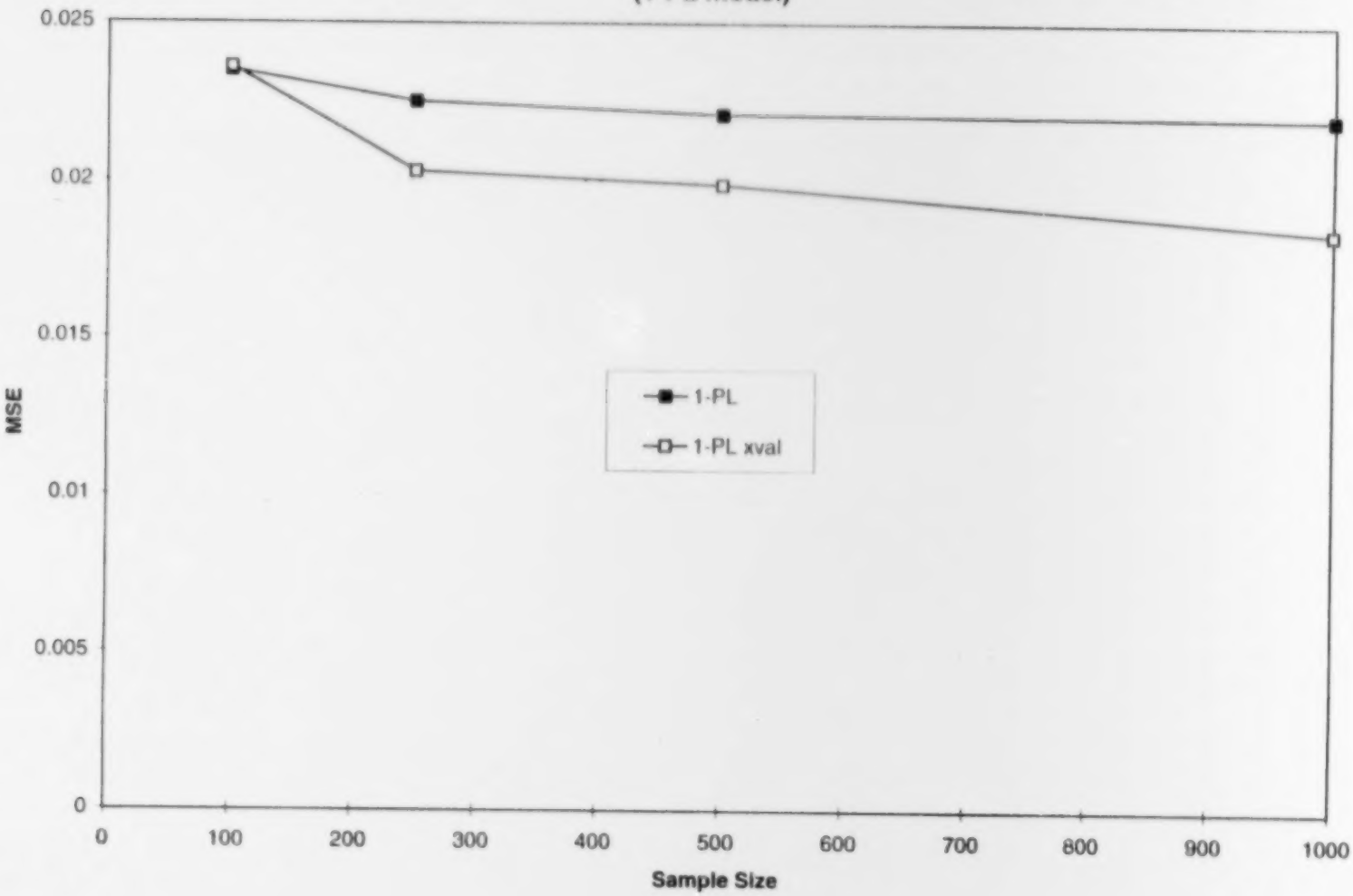


Figure 10 - Mean Squared Error Between Estimated and Expected Response Probabilities (2-PL models)

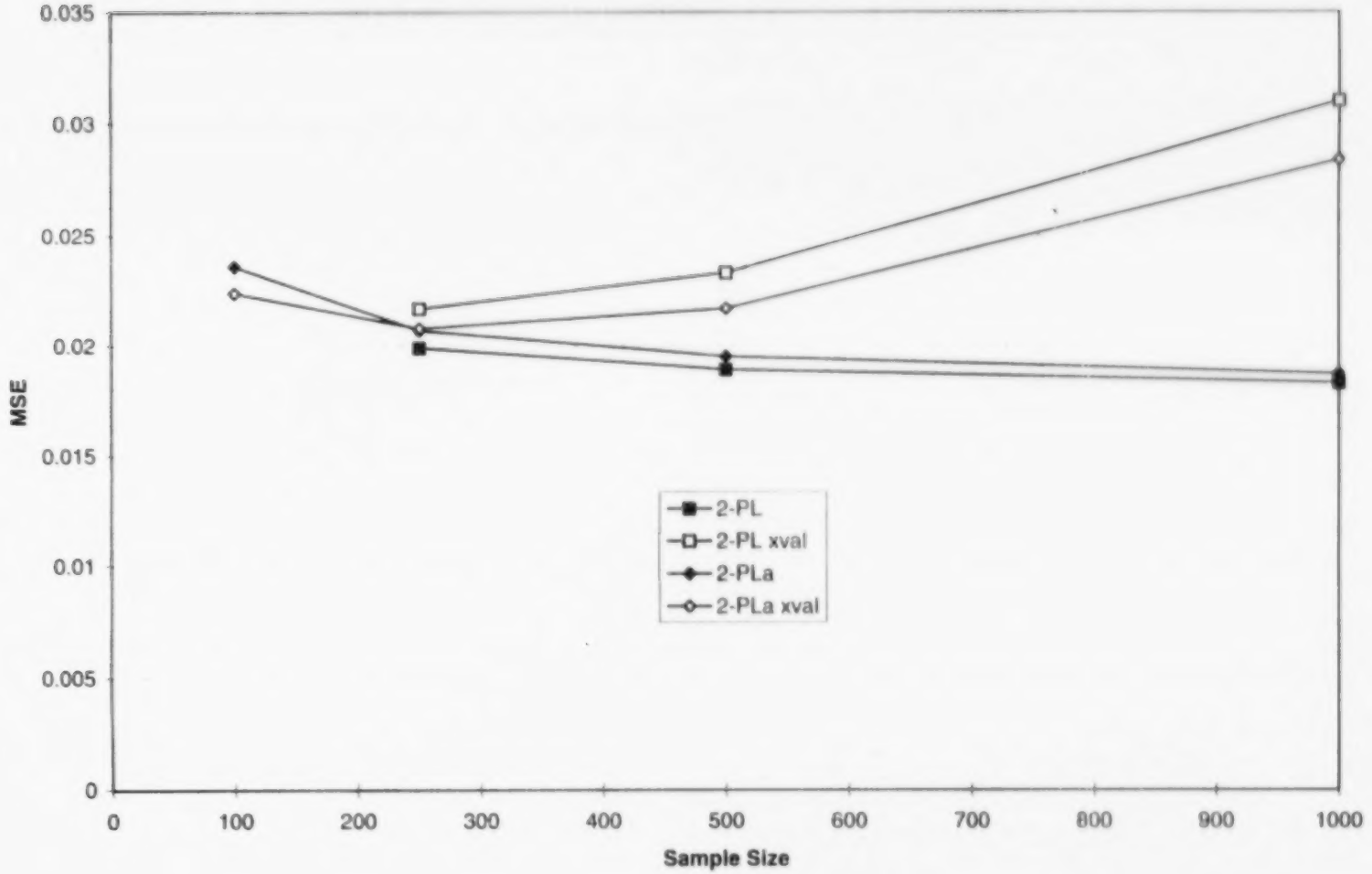


Figure 11 - Mean Squared Error Between Estimated and Expected Response Probabilities (3-PL models)

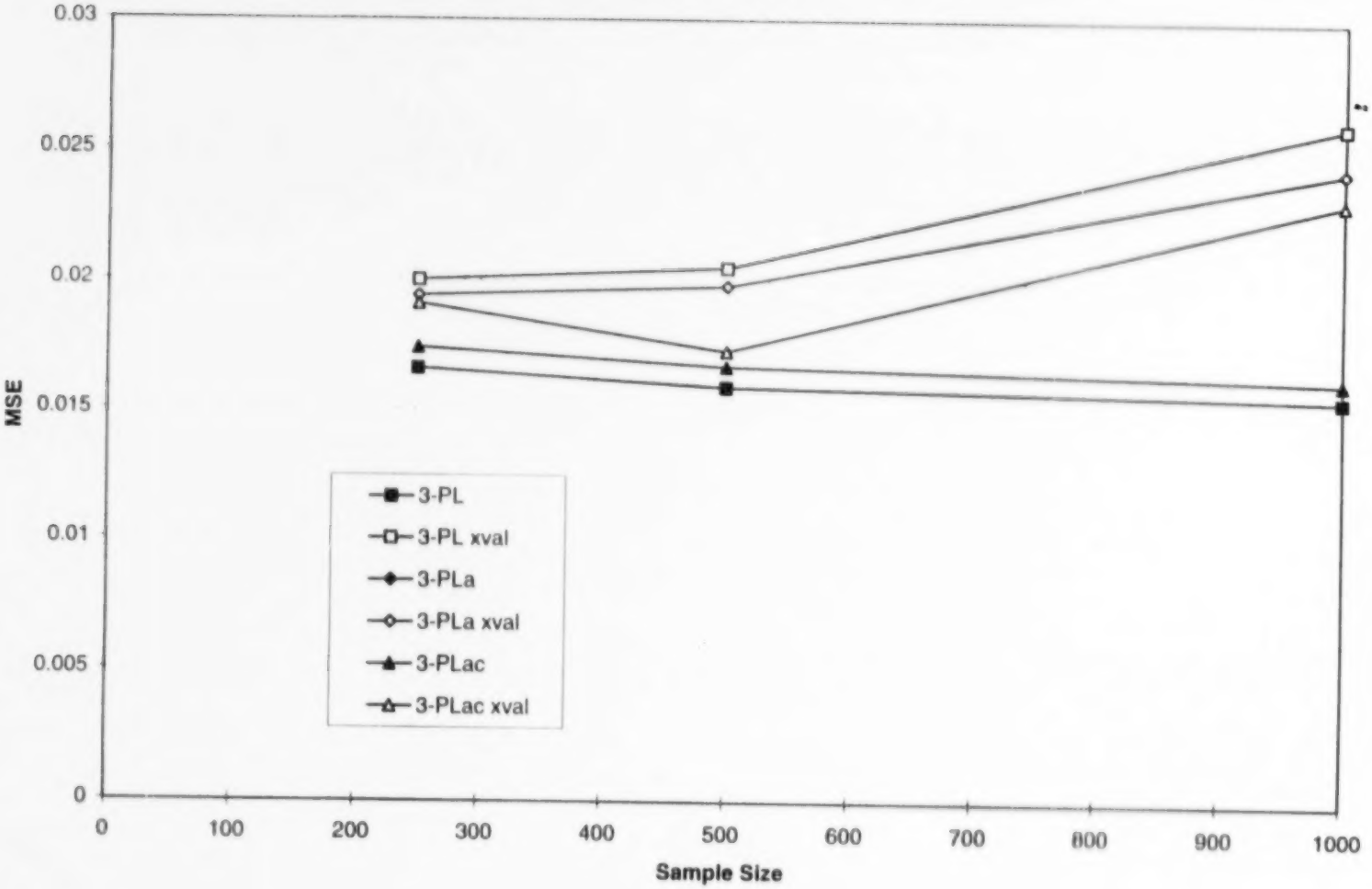


Figure 12 - Root Mean Squared Error Between Estimated and Expected Number Correct Scores (1-PL model)

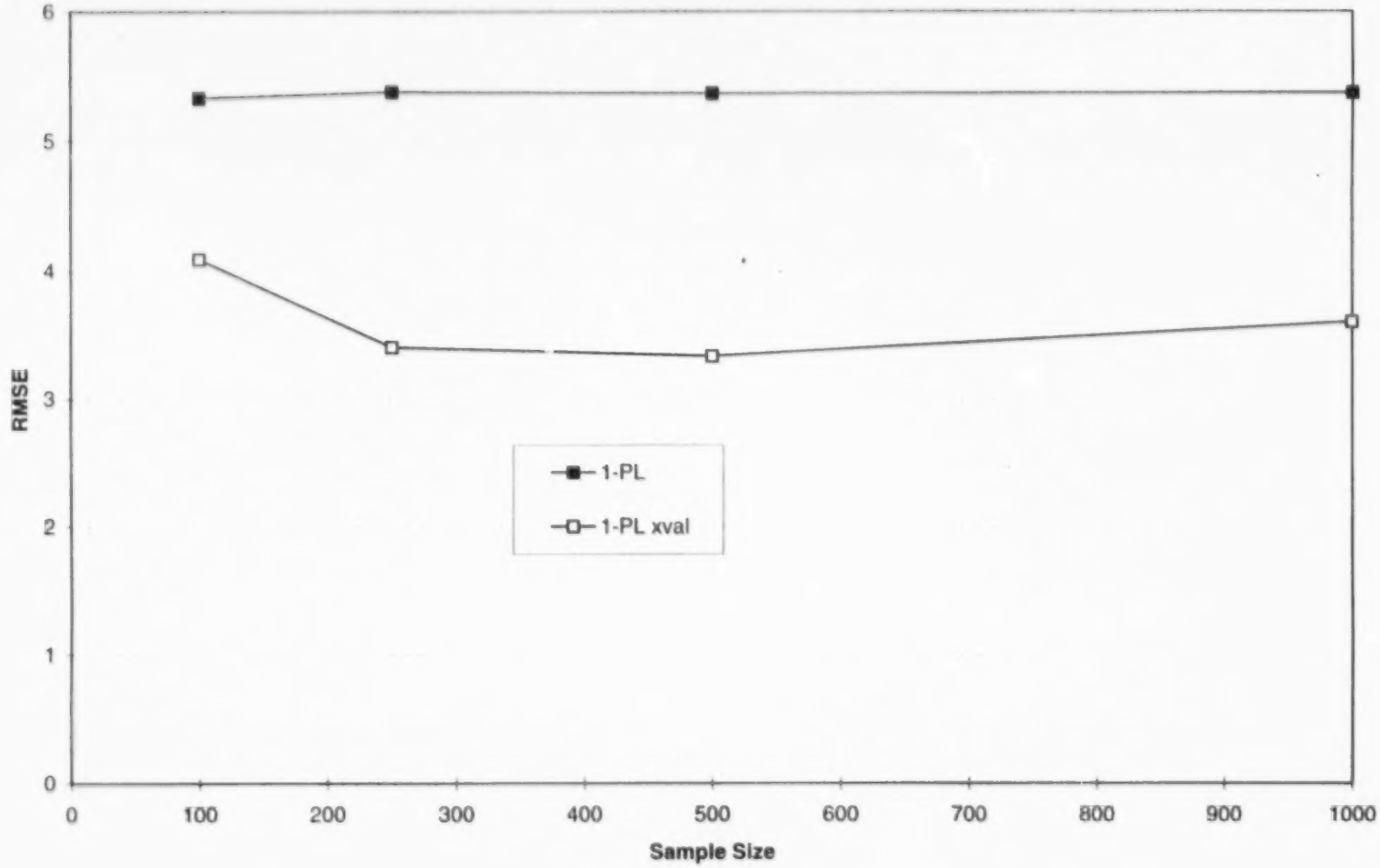


Figure 13 - Root Mean Squared Error Between Estimated and Expected Number Correct Scores (2-PL models)

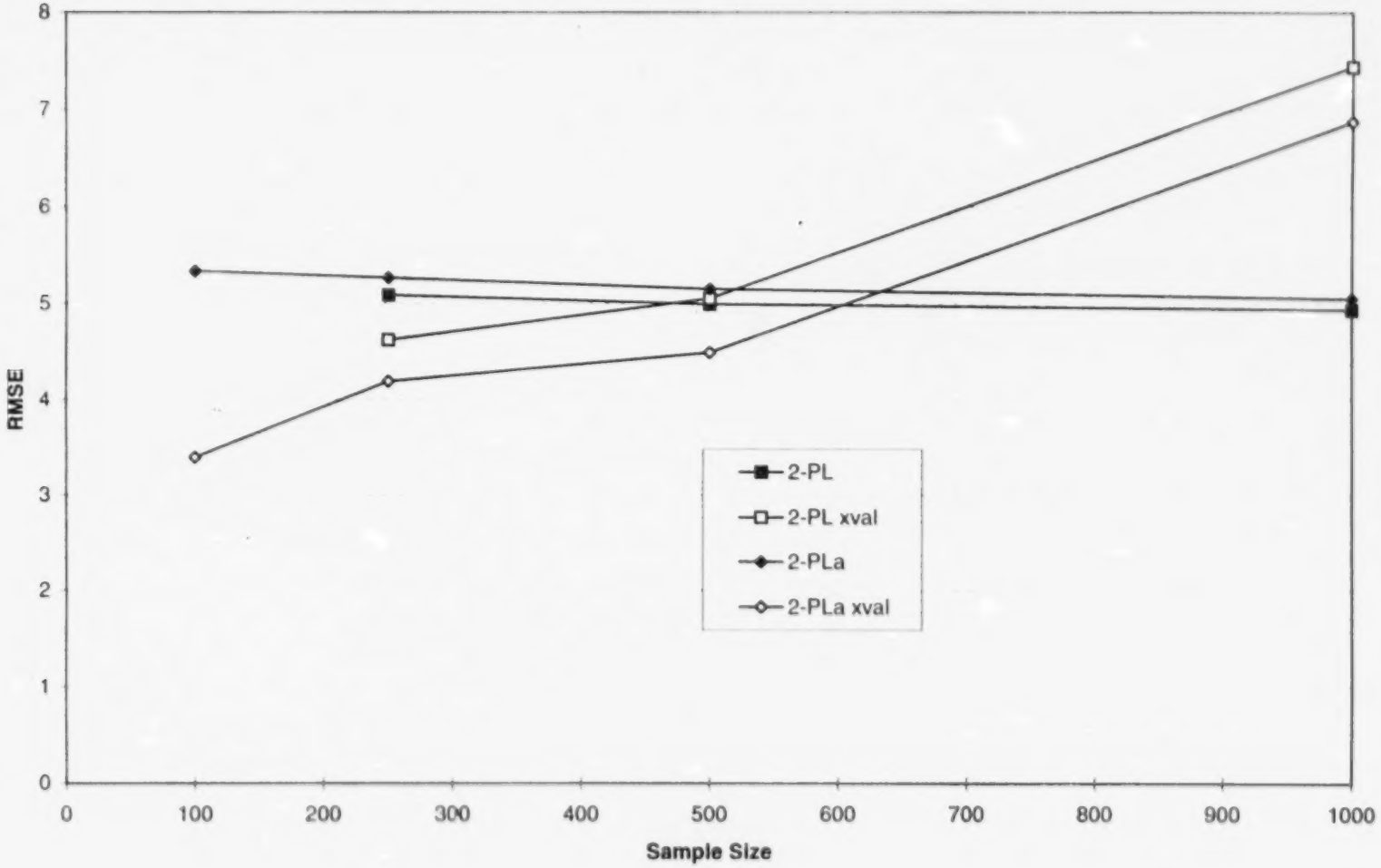


Figure 14 - Root Mean Squared Error Between Estimated and Expected Number Correct Scores (3-PL models)

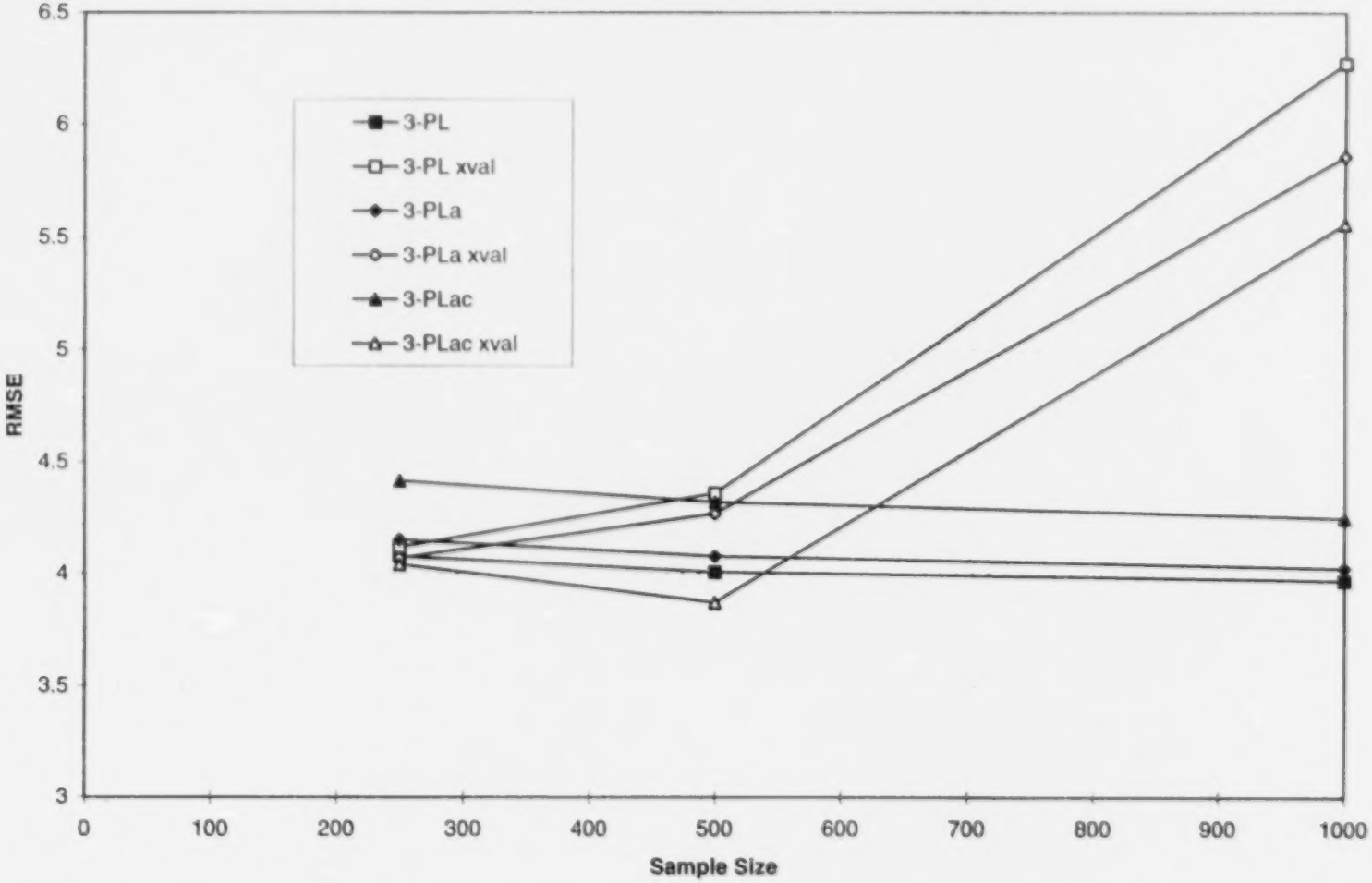


Figure 15 - Spearman Rank Correlation Between Estimated and Expected Number Correct Scores (1-PL model)

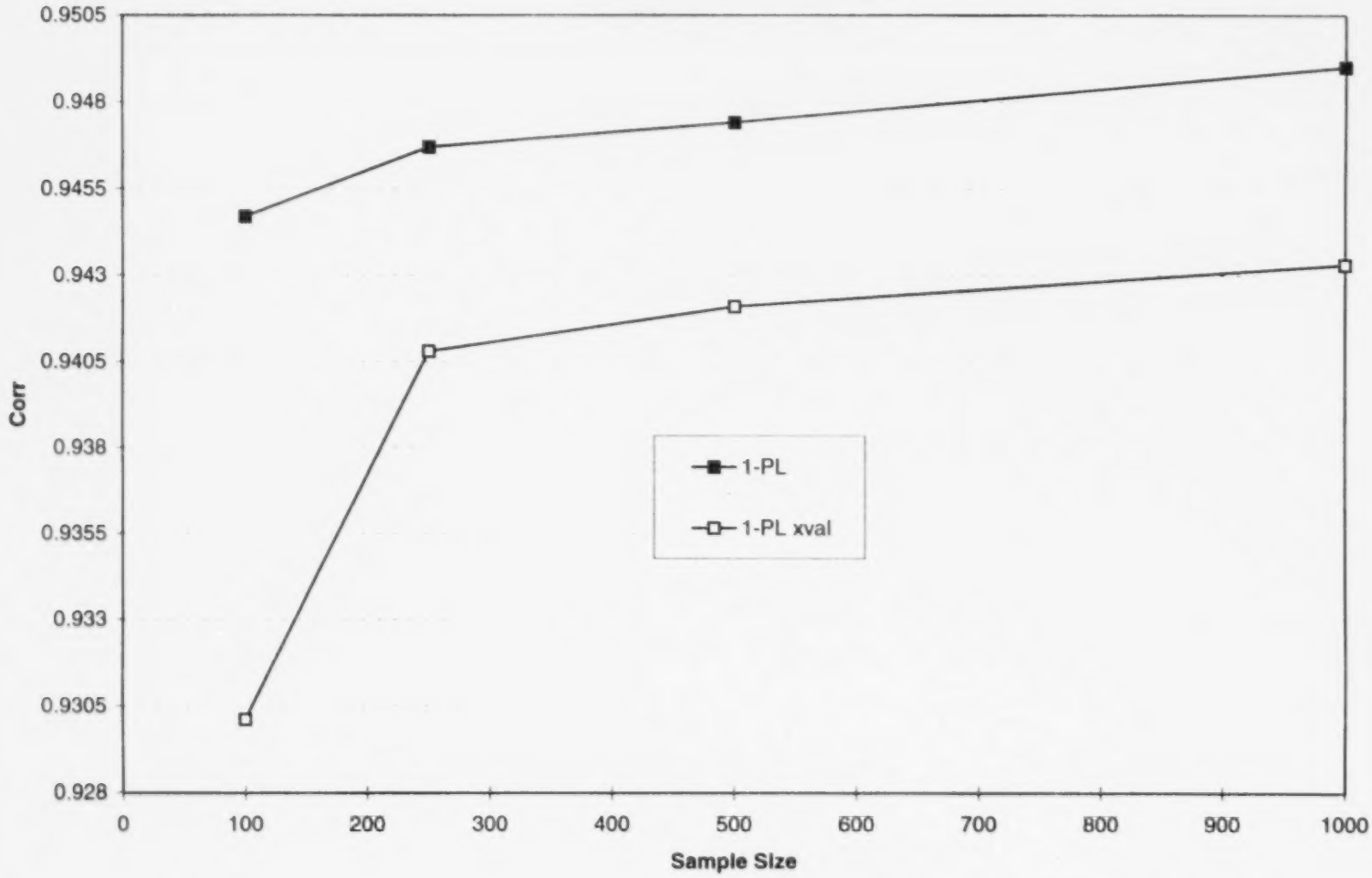


Figure 16 - Spearman Rank Correlation Between Estimated and Expected Number Correct Scores (2-PL models)

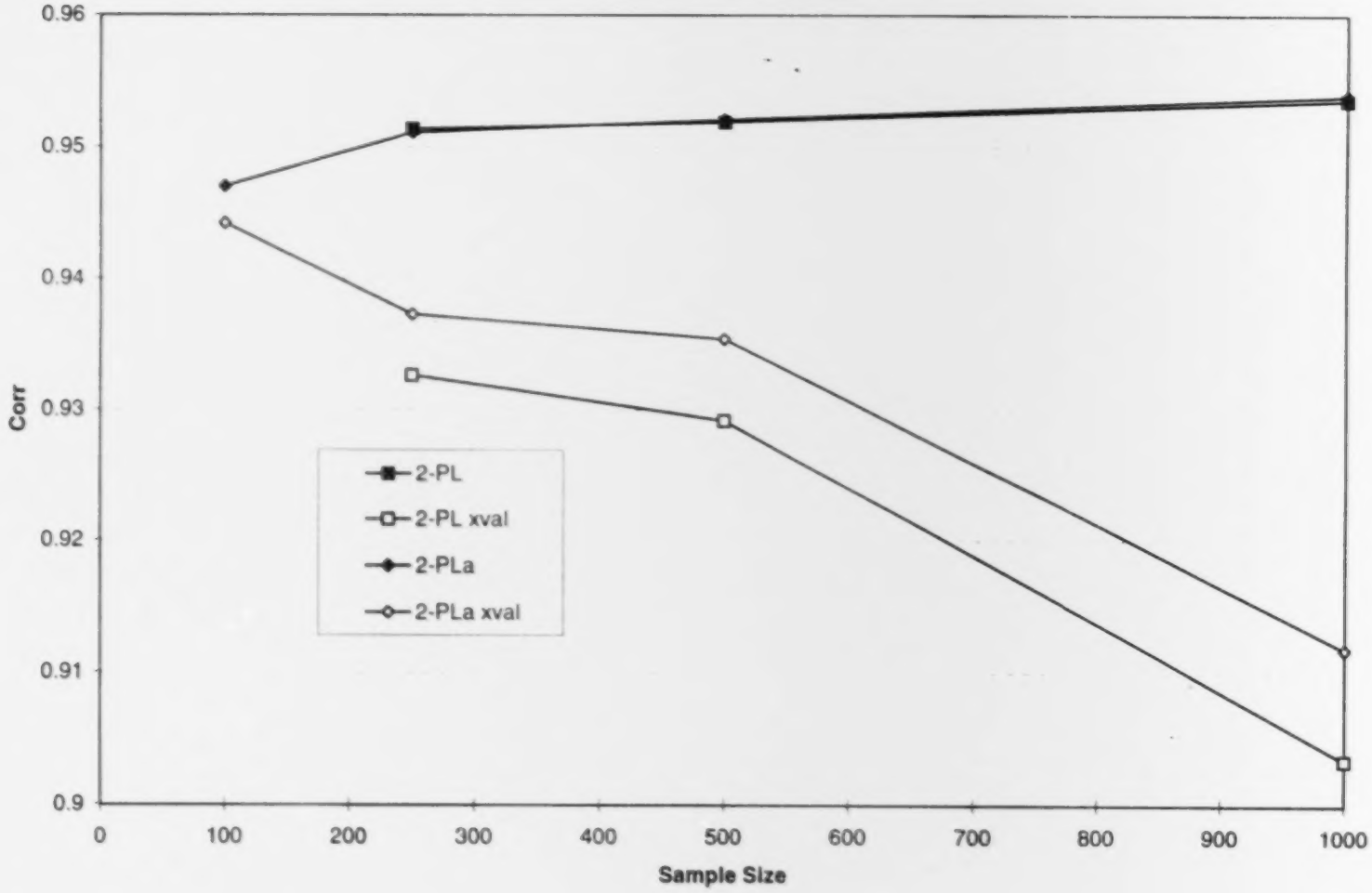


Figure 17 - Spearman Rank Correlation Between Estimated and Expected Number Correct Scores (3-PL models)